Om kvantilregresjon

Masteroppgave i matematisk statistikk
(30 studiepoengs omfang)

Jon Harald Leknes

Matematisk institutt
Universitetet i Bergen

Desember 2008

Sammendrag

Vi skal i denne oppgaven beskrive teorien bak kvantilregresjon. I klassisk regresjonsanalyse bruker man minste kvadraters metode, hvor man beregner betingede gjennomsnittsestimater av forventningen til responsvariabelen. Vi skal vise hvordan man i stedet kan minimere absoluttavstand, vektet i forhold til hvilken kvantil man ønsker å se på, slik at vi kan regne ut betingede kvantilestimater. Gjennom kvantilregresjon kan vil altså se på hele fordelingen til det vi ønsker å estimere.

I tillegg vil den betingede sentralkvantilen (medianestimatet) bli lansert som et alternativ til den betingede gjennomsnittsverdien som sentralestimat, og vi vil se på hvorfor medianestimatet kan være å foretrekke i visse situasjoner.

Til slutt skal vi se på kvantil autoregresjon, som er bruk av kvantilregresjonsmetoder for autoregressive modeller i tidsrekkeanalyse. Her går de klassiske metodene også ut på å minimere kvadratavstand. Vi vil se at vi i forbindelse med kvantil autoregresjon vil møte utfordringer forskerne ikke er enige om løsningen på ennå.

Vi vil underveis benytte oss av eksempler for å belyse teorien som blir beskrevet.

Innhold

Forord

Jeg har aldri vært glad i å ha for god tid på meg når jeg skal få noe gjort, så da jeg ble opplyst om at man kunne velge å skrive masteroppgave på 30 studiepoeng, altså ett semester, så var valget enkelt. Jeg vil takke min veileder, Dag Tjøstheim, for å ha foreslått temaet for oppgaven, som jeg utelukkende har hatt positive erfaringer å jobbe med. Og ikke minst en stor takk til Dag for samarbeidet underveis, som har vært til stor inspirasjon og motivasjon for meg. Mange takk også til min klassevenninne gjennom tolv år, Marianne Tangen Bråthen, for hjelp med korrekturlesing.

Oppgavens oppbygning

Det kan være en utfordring å sette seg inn i teorien for kvantilregresjon på kort tid ved å bla i litteraturen som finnes om temaet. Hensikten med denne oppgaven er derfor at leseren på en lett forståelig måte skal få en innføring i det mest sentrale stoffet, hva som er motivasjonen bak, hvordan det fungerer og hvorfor man er så avhengig av statistisk programvare. Jeg har også ønsket å få fram kontrasten mellom metodikk for ordinær minste kvadraters analyse og kvantil analyse. Jeg har derfor tatt med utledninger for gjennomsnitt og ordinære regresjonsestimater som vanligvis sløyfes i masteroppgaver, i den hensikt å nettopp illustrere disse forskjellene. For relevant stoff om kvantilregresjon utover det grunnleggende som får plass i denne oppgaven henvises det til [Koenker 2005], en bok om kvantilregresjon skrevet av Roger Koenker, som er den mest sentrale personen bak utviklingen av teorien om dette temaet.

Kapittel 1 er en innledning som tar for seg historikken til kvantilregresjon, og motivasjonen for innføringen av denne teorien. I kapittel 2 definerer vi gjennomsnitt, median og kvantiler, samt andre begreper tilknyttet dette. I kapittel 3 ser vi på hvordan man kan definere gjennomsnitt, median og kvantiler som minimaliseringsproblem. Det er slike minimaliseringsproblem man bruker i regresjonsanalyse, som vi skal se på i kapittel 4. Vi vil der starte med klassisk minste kvadraters regresjon, og videre utlede teorien for medianregresjon og kvantilregresjon, og underveis illustrere med eksempler. I tillegg vil vi vise hvordan vi bruker den statistiske programvaren `R` i forbindelse med kvantilregresjon, og til slutt i kapittel 4, i seksjon 4.6, ser vi på noen eksempler på et større datasett. Teorien som blir tatt opp i kapittel 3 og 4 er i hovedsak hentet fra nevnte [Koenker 2005] og [Hao og Naiman 2007], sistnevnte et hefte fra 2007 som tar for seg anvendelse av kvantilregresjon i sosiale studier, hvor skjeve fordelinger spiller en sentral rolle. I kapittel 5 ser vi på tidsrekkeanalyse, hvor vi kort beskriver den klassiske teorien for denne, og viser hvordan kvantilregresjonsmetoder kan benyttes. Dette kapitlet er i hovedsak basert på [Koenker og Xiao 2006], en tidsskriftartikkel fra 2006 som introduserer bruk av kvantilregresjonsmetoder på tidsrekkeanalyse. Med andre ord er dette ganske nytt, og vi vil se på problemer med denne teorien hvis løsning forskerne ennå ikke er enige om.

1 Innledning

I statistikk er vi som regel interessert i mer enn bare gjennomsnittsverdier. Francis Galton skrev i 1889 at det å kun se på gjennomsnittsverdier i statistikk er altfor snevert. Siden den gang har det vært en stor utvikling i statistikk. Man har vært interessert i hvordan observasjoner fra en gitt fordeling varierer i verdi, skjevhet, hvor de forskjellige kvantilene ligger, hypotesetesting og konfidensintervall, og teori for alt dette har vært etablert en god stund. I dag blir begrepet "variasjonsbredde", differansen mellom største og minste observasjon i en endelig mengde, innført allerede i ungdomsskolen.

I regresjonsanalyse har man lenge vært på etterskudd; her har man tradisjonelt bare vært opptatt av å finne betingede gjennomsnittsestimater. I for eksempel en enkel, lineær regresjonsmodell får vi med de tradisjonelle metodene en regresjonslinje som viser hvilken verdi, gjennomsnittlig, responsvariabelen vil få når forklaringsvariabelen er gitt, basert på de observasjonene som dannet grunnlaget for analysen (se kapittel 4). Dette gjennomsnittsprinsippet, som går ut på å minimere kvadratavstand (se seksjon 3.1), gjelder for all vanlig regresjonsanalyse.

Det har opp gjennom tidene vært kritikk fra statistikere angående mangel på generelle metoder som tar for seg mer enn bare gjennomsnitt når det gjelder regresjonsanalyse.

Mosteller og Tukey skrev i 1977 at regresjonsanalyse bare gir et bilde av gjennomsnittsverdier, og at man kan gå videre ved å regne ut regresjonskurver for forskjellige kvantiler (se seksjon 2.3 for definisjon av hva en kvantil er) for å få et mer fullstendig bilde. De antyder videre at den vanlige regresjonskurven alene gir et like lite komplett bilde av sammenhengen mellom to fordelinger som kun gjennomsnittet alene gir for en enkel fordeling.

Ville det ikke kunne være nyttig med et system hvor vi, for eksempel for enkel lineær regresjon, kan finne en familie regresjonslinjer for alle ønskelige kvantiler? Og se om det er forskjeller i utviklingen ved de ulike kvantilene, det vil si se på hele den betingede fordelingen? Og så generalisere dette til å gjelde alle former for regresjonsanalyse? Det er dette som er motivasjonen bak kvantilregresjon, introdusert av Koenker og Bassett i 1978.

Et forløp til en idé for kvantilregresjon ble lansert allerede i 1760-årene av den allsidige kroatiske vitenskapsmannen Rudjer Jusip Boscovich. Konseptet medianregresjon går ut på, i stedet for å se på gjennomsnitt ved å minimere kvadratavstand, å omformulere slik at man utfører regresjonsanalyse hvor man lager medianestimater til regresjonskoeffisientene, ved å minimere absoluttavstand (se seksjon 3.2). I og med

at medianen er det samme som 0.50-kvantilen er veien kort for å generalisere, slik at man kan utføre regresjon basert på alle ønskelige kvantiler, dette ved en type vektet absoluttavstand (se seksjon 3.3). Boscovich utførte omformuleringen fra gjennomsnitt til median angående regresjonskoeffisientene i en enkel, lineær regresjonsmodell. Men verken han eller Edgeworth, som drøyt 100 år senere forsøkte å generalisere dette til en multippel regresjonsmodell, lyktes i å finne ut hvordan man i praksis burde gå fram for å utføre de nødvendige beregninger som må til for å tilpasse en modell til observasjonsmengder. Teorien for lineær programmering (se side 46) som måtte til for å løse problemet ble ikke utviklet før i 1940-årene, men selv da var tilsynelatende enkle problemer tidkrevende å løse. Noen tiår senere utviklet datamaskiner seg til å bli et vanlig verktøy å utføre matematiske beregninger på, og dette er en viktig grunn til at kvantilregresjon først har slått til de senere år. Man har fått utviklet gode programsystem for statistikk, hvor man blant annet har implementert de nokså kompliserte metodene for kvantilregresjon, slik at datamaskiner raskt kan gjøre utregninger og tilpasse modeller for oss.

Motivasjonen til Edgeworth var at han oppdaget at et medianestimat i mange situasjoner er mer robust enn en estimator basert på minimert kvadratavstand, som var vel etablert på den tiden og som fungerer best under ideelle forutsetninger; når vi har normalfordeling hvor variansen er uavhengig av forklaringsvariabelen. Dette er en annen, viktig motivasjonsfaktor for kvantilregresjon, nemlig å erstatte gjennomsnitt med det generelt mer robuste målet median som sentralmål.

Kvantilregresjon kan være et nyttig verktøy innenfor mange retninger i statistikk. I for eksempel tidsrekkeanalyse benytter man seg tradisjonelt også av minste kvadraters metode for å finne estimater. Hvorfor bare estimere for eksempel hvordan tidsrekker gjennomsnittlig vil utvikle seg – hva med hvordan de forskjellige kvantilene utvikler seg, eller hele fordelingen til prediktorvariabelen? Dette kan gjøres ved å bruke kvantilregresjonsmetoder, hvor man minimerer vektet absoluttavstand i stedet for kvadratavstand.

2 Gjennomsnitt, median og kvantiler

Vi vil etter hvert få mye bruk for begrepene *gjennomsnitt*, *median* og *kvantiler*, og for å ha et utgangspunkt å gå videre fra vil vi derfor definere disse til å begynne med, selv om dette er ytterst velkjente størrelser, se seksjon 2.1 til 2.3. Disse begrepene utgjør kjernen i teorien som belyses utover i denne oppgaven, og en grundig forståelse av dem er derfor nødvendig. I dette kapitlet skal vi i tillegg definere en del andre begreper tilknyttet dette, hvor [Casella og Berger 2002] er primærkilden. Lesere som ikke er interessert i oppfriskning av slikt elementært stoff kan hoppe over dette kapitlet.

Gjennomsnitt er det mest brukte sentralmålet vi har for data. Forventningsverdien, som er det mest brukte sentralmålet for sannsynlighetsfordelinger, bygger på gjennomsnittsprinsippet. Dette prinsippet er det samme som anvendes i vanlig regresjonsanalyse.

Median er et alternativt sentralmål, og vi skal se på hva som er forskjellen mellom gjennomsnitt og median, og i hvilke situasjoner forskjellen er signifikant. Kvantiler er en generalisering av median, der vi i stedet for å se på sentralkvantilen (som er det samme som medianen) kan rette oppmerksomheten mot et hvilket som helst annet sted i fordelingen. Hva dette vil si kommer vi nærmere inn på når vi definerer kvantiler, se seksjon 2.3.

Vi skal senere se at medianprinsippet er det som brukes i medianregresjon, og dette kan så generaliseres til kvantilregresjon, slik at vi kan lage regresjonskurver som representerer estimater av responsfordelingen på ulike steder og ikke bare sentralt.

Sannsynlighetstetthet, kumulativ fordelingsfunksjon og punktsannsynlighet

Vi vil først definere begrepene *sannsynlighetstetthet*, *kumulativ fordelingsfunksjon*, og *punktsannsynlighet*, som vi får bruk for i noen av definisjonene for gjennomsnitt, median og kvantiler. For detaljer, se [Casella og Berger 2002] side 27-36.

Det kontinuerlige tilfellet

En stokastisk variabel X med kontinuerlig sannsynlighetsfordeling har sannsynlighetstetthet $f(x)$, der $f(x)dx$ er sannsynligheten for at et utfall er i intervallet $[x, x+dx]$. Vi må ha $f(x) \geq 0$, siden sannsynlighet må være i intervallet $[0, 1]$. Vi har videre betingelsen $\int_{-\infty}^{\infty} f(x)dx = 1$, som betyr at total sannsynlighet må være lik 1.

Den kumulative fordelingsfunksjonen er definert som $F(x) = \int_{-\infty}^{x} f(t)dt$ eller $F(x) = P(X \leq x)$, der $P(X \leq x)$ er sannsynligheten for at variabelen X er mindre eller lik verdien x. Siden $f(x) \geq 0$ følger det at F er en monotont voksende funksjon, der $\lim_{x\to-\infty} = 0$ og $\lim_{x\to\infty} F(x) = 1$, og vi ser at $F'(x) = f(x)$.

Det diskrete tilfellet
En stokastisk variabel med diskret sannsynlighetsfordeling har punktsannsynlighet $P(X = x) \geq 0$, der $P(X = x)$ er sannsynligheten for at den stokastiske variabelen X er lik verdien x. Vi har videre betingelsen $\sum_{\forall x} P(X = x) = 1$, som betyr at total sannsynlighet må være lik 1.

Den kumulative fordelingsfunksjonen er definert som $F(x) = P(X \leq x)$. Siden $P(X = x) \geq 0$ følger det at F er en monotont voksende funksjon, hvis verdi må befinne seg i intervallet $[0, 1]$. Dette på grunn av at sannsynlighet også må være i intervallet $[0, 1]$.

2.1 Gjennomsnitt

2.1.1 Definisjon av empirisk gjennomsnitt

Gjennomsnittet $\overline{x}$ av en endelig mengde observasjoner X er definert til å være summen av hvert enkelt element delt på antall elementer:

$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Vi skal videre forholde oss til en fiktiv sortert mengde U av sju observasjoner,

$u_i : \{4, 5, 6, 8, 9, 11, 13\}$.

For U har vi:

$\overline{u} = \frac{1}{7} \sum_{i=1}^{7} u_i = 8$.

Intuitivt virker gjennomsnitt å være et fornuftig mål på sentralitet, men vi skal senere se at dette sentralmålet har sine svakheter i visse situasjoner (se for eksempel seksjon 2.4 på side 13).

2.1.2 Definisjon av forventning

Gjennomsnitt som definert over er en empirisk størrelse. Vi bruker betegnelsen *forventning* om det vi kan kalle teoretisk gjennomsnitt. Vi har følgende definisjon for forventningsverdien til en kontinuerlig stokastisk variabel med sannsynlighetstetthet $f(x)$:

$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$.

For diskrete fordelinger med punktsannsynlighet $P(X = x)$ har vi:

$\mathrm{E}(X) = \sum_{\forall x} xP(X = x)$.

Vi kan vise sammenhengen mellom forventningen og gjennomsnittet ved å for eksempel se på en vanlig, sekssidet terning. Er terningen rettferdig har vi at $P(X = x_i) = \frac{1}{6}, x_i : \{1, 2, 3, 4, 5, 6\}$. Forventningsverdien om vi kaster én gang blir $\mathrm{E}(X) = \frac{1}{6}\sum_{i=1}^{6} x_i = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3.5$, som er det samme som gjennomsnittet av mengden $\{1, 2, 3, 4, 5, 6\}$, som var utfallsrommet i dette tilfellet.

Hvis forskjellige observasjoner har ulik sannsynlighet, så viser formlene at vi vil vekte observasjonene med tilhørende sannsynlighet, og det faktum at vektene totalt summeres til 1 gjør at forventning blir som et vektet gjennomsnitt. Ellers har vi store talls sterke lov (se [Casella og Berger 2002] side 235), som sier at hvis $X_1, X_2, ..., X_n$ er uavhengig, identisk fordelte stokastiske variable med endelig varians, så vil gjennomsnittet $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ konvergere mot forventningsverdien $\mathrm{E}(X)$ når $n \to \infty$.

Fordeler og ulemper med gjennomsnitt som sentralmål
Fordelen med gjennomsnitt er at størrelsen på alle observasjonene spiller inn. Dette kan imidlertid i noen situasjoner føre til en ulempe, for dersom noen få observasjoner for eksempel er unormalt store relativt til resten[1], så vil disse observasjonene prege resultatet i stor grad, og gjennomsnittet vil da ikke gi noe godt bilde på en typisk observasjon. Se seksjon 2.4 for et eksempel på akkurat dette.

2.2 Median

2.2.1 Definisjon av teoretisk median

Medianen, M, i en sannsynlighetsfordeling, er verdien som deler sannsynlighetsmassen i to like store deler (i den grad det er mulig for en diskret fordeling). Vi har generelt $P(X \leq M_X) \geq 0.50$ og $P(X \geq M_X) \geq 0.50$. I det kontinuerlige tilfellet kan dette skrives som $P(X \leq M_X) = \int_{-\infty}^{M_X} f(x)dx = P(X \geq M_X) = \int_{M_X}^{\infty} f(x)dx = 0.50$. Sannsynligheten er altså her 0.50 for å få en verdi som er mindre enn medianen, og dermed 0.50 for å få en verdi som er større enn medianen.

For symmetriske fordelinger har vi $M_X = \mathrm{E}(X)$, siden forventningen da ligger i midten der sannsynlighetsmassen deles i to like store deler. Men straks vi blir stilt overfor skjeve fordelinger vil vi se at forskjellen mellom disse to sentralmålene kan være betydelig (se eksempelet i seksjon 2.6 på side 17).

[1] Vi vil videre benytte omgrepene *uteligger* eller *uteliggende observasjon* om slike observasjoner.

2.2.2 Definisjon av empirisk median

Den empiriske medianen, $\hat{M}$, i en endelig mengde av observasjoner, er definert til å være den midterste observasjonen når mengden er sortert[2]. Har vi n sorterte observasjoner og n er odde, er $\hat{M}$ lik observasjon nummer $\frac{n+1}{2}$, det vil si at vi har like mange observasjoner på hver side av $\hat{M}$. Er n jevn, er $\hat{M}$ definert til å være lik gjennomsnittet av de to observasjonene i midten, observasjon nummer $\frac{n}{2}$ og observasjon nummer $(\frac{n}{2}+1)$ når mengden er sortert. Vi har altså også her like mange observasjoner på hver side av $\hat{M}$.

For mengden U definert på side 9 har vi $\hat{M}_U = 8$, altså har vi $\hat{M}_U = \overline{u}$ her, men akkurat som i det teoretiske tilfellet skal vi se at disse to empiriske sentralmålene i visse situasjoner kan gi svært ulikt resultat (se seksjon 2.4 på side 13).

Fordeler og ulemper med median som sentralmål
Fordelen med den empiriske medianen er at den ikke tar hensyn til uteliggere overhodet. Se seksjon 2.4 for et eksempel på dette. Ulempen er at den kun tar hensyn til det som skjer i midten, og dersom observasjonene rundt varierer på ulik måte på hver side av den empiriske medianen, så spiller ikke dette inn på medianestimatet.

2.3 Kvantiler

Kvantiler er en generalisering av median. Vi definerer $Q^{(\tau)}$ til å være τ-kvantilen, der τ inngår i intervallet $[0, 1]$. Vi skal se at τ står i sammenheng med definisjonen av sannsynlighet, som også ligger i intervallet $[0, 1]$. Denne definisjonen av τ antas kjent videre. Medianen, M eller $Q^{(0.50)}$, er 0.50-kvantilen.

2.3.1 Definisjon av kvantiler i teorien

Vi forholder oss til en stokastisk variabel X. Vi har $Q_X^{(\tau)} = \inf\{x : F(x) \geq \tau\}$, altså er τ-kvantilen $Q_X^{(\tau)}$ det minste settet av x-verdier som er slik at vi får $F(x) \geq \tau$. Alternativt har vi formuleringen $P(X \leq Q_X^{(\tau)}) \geq \tau$ og $P(X \geq Q_X^{(\tau)}) \geq 1 - \tau$, som er en generalisering av definisjonen for medianen på forrige side.

I det kontinuerlige tilfellet kan dette skrives som $P(X \leq Q_X^{(\tau)}) = \int_{-\infty}^{Q_X^{(\tau)}} f(x)dx = \tau$ og $P(X \geq Q_X^{(\tau)}) = \int_{Q_X^{(\tau)}}^{\infty} f(x)dx = 1 - \tau$. Sannsynligheten er altså her τ for å få en verdi som er mindre enn $Q_X^{(\tau)}$, og dermed $1-\tau$ for å få en verdi som er større enn $Q_X^{(\tau)}$. Vi ser at medianen er et spesialtilfelle, da $Q_X^{(\tau)} = Q_X^{(0.50)}$ gir den samme definisjonen som for teoretisk median på forrige side, altså har vi generelt $Q^{(0.50)} = M$.

[2]At en mengde er sortert betyr at vi har plassert alle elementene etter hverandre i stigende rekkefølge.

2.3.2 Definisjon av empiriske kvantiler

Vi har i teorien i det kontinuerlige tilfellet at 100τ % av observasjonene skal være mindre eller lik den teoretiske kvantilen $Q^{(\tau)}$ og $100(1-\tau)$ % skal være større eller lik. Som i det diskrete tilfellet i teorien vil dette ofte ikke la seg gjøre empirisk, siden vi i observert mengde må ha et endelig antall observasjoner. Vår endelige definisjon vil være basert på definisjonen i teorien på forrige side, men først skal vi gjennom et eksempel beskrive noen alternative måter empiriske kvantiler kan defineres på.

For den sorterte mengden U fra side 9 skal vi finne $\hat{Q}_U^{(0.20)}$. Vi kan definere flere måter å gjøre dette på. Vi har bare sju observasjoner, og dermed kan vi si at $\hat{Q}_U^{(0.20)}$ er lik observasjon nummer $7 \cdot 0.20 = 1.4$. Velger vi å bruke vekting av observasjoner regner vi ut at $\hat{Q}_U^{(0.20)} = u_1 + 0.4(u_2 - u_1) = 4.4$. I [Casella og Berger 2002] på side 227 er det utledet en definisjon hvor man velger den nærmeste observasjonen, og vi vil dermed få $\hat{Q}_U^{(0.20)} = u_1 = 4$. Man har her en spesiell definisjon for $\tau = 0.50$, nemlig samme definisjon som for empirisk median, altså generelt $\hat{Q}^{(0.50)} = \hat{M}$. Vi vil se under at om vi ikke innfører denne spesielle definisjonen, så får vi generelt $\hat{Q}^{(0.50)} \neq \hat{M}$ når vi har jevnt antall observasjoner.

Bruker vi samme type definisjon som vi beskrev for kvantiler i teorien, som er den definisjonen som vil gjelde videre i denne oppgaven, får vi at vi må ha det minste settet av mengden U (når den er sortert) som er slik at vi får $F_7(u) \geq \tau$, der $F_7(u)$ er den empiriske fordelingsfunksjonen til settet U, definert som $F_7(u) = \left\{ \frac{\#u_i : u_i \leq u}{n} \right\}$. Tallet 7 kommer av at vi har $n = 7$ i eksempelet vårt. Vi må her ha $F_7(u) \geq 0.20$. Vi lar u her være et hvilket som helst reelt tall, og den minste u-verdien som oppfyller kravet $F_7(u) \geq 0.20$ vil bli $\hat{Q}_U^{(0.20)}$, kvantilestimatet vi er ute etter. For $u = u_1$ får vi $F_n(u) = \frac{1}{7} < 0.20$ mens for $u = u_2$ får vi $F_n(u) = \frac{2}{7} \geq 0.20$. Vi må altså runde av oppover i stedet for til nærmeste tall, slik at vi får $\hat{Q}_U^{(0.20)} = u_2 = 5$.

Vi ser her at vi ved denne definisjonen generelt vil få $\hat{Q}_X^{(0.50)} = x_{\frac{n}{2}}$ når X er et sortert sett som inneholder jevnt antall observasjoner. Fra side 11 hadde vi for jevnt antall observasjoner den spesielle definisjonen $\hat{M}_X = \frac{1}{2}(x_{\frac{n}{2}} + x_{(\frac{n}{2}+1)})$, slik at $\hat{M}_X$ skiller seg noe fra $\hat{Q}_X^{(0.50)}$ i denne situasjonen. Vi vil senere se i kvantilregresjon (se Eksempel 4.3 fra side 41) at man i teorien som er utviklet der ikke definerer den empiriske medianen som noe annet enn den empiriske 0.50-kvantilen, slik at det er denne definisjonen som vil gjelde videre.

2.3.3 Egenskaper til kvantiler

Empiriske kvantiler lar seg ikke påvirke av uteliggere, av samme grunn som for spesialtilfellet empirisk median (se seksjon 2.4). Hvor mye større observasjonene som er større enn et kvantilestimat er spiller ikke inn (tilsvarende gjelder for observasjoner

som er mindre). Lager vi oss en mengde med flere kvantiler kan vi få god oversikt over fordelingen til en variabel, for eksempel de 19 kvantilene $\{Q^{(0.05)}, Q^{(0.10)}, ..., Q^{(0.95)}\}$. Vi vil senere se at det samme vil gjelde for kvantilregresjon kontra ordinær, gjennomsnittbasert regresjon (se kapittel 4). 19 regresjonskurver som representerer de nevnte kvantilene gir et mye bedre bilde av sammenhengen mellom en responsvariabel og en forklaringsvariabel enn bare én kurve som representerer gjennomsnittet.

I tillegg har vi at om vi har en monoton funksjon h, og hvis $Q_X^{(\tau)}$ er τ-kvantilen av X, så vil $h(Q_X^{(\tau)})$ være τ-kvantilen til $h(X)$. Dette gjelder også for empiriske kvantiler. En lignende regel gjelder ikke for forventningsverdier, hvor vi generelt har $\mathrm{E}[h(X)] \neq h[\mathrm{E}(X)]$, med enkelte unntak, for eksempel når h er en lineær funksjon.

2.4 Robusthet for uteliggeres innflytelse

Vi skal nå se hva som kan skje når vi har en uteligger i en observasjonsmengde. Vi definerer $U^* = U \bigcup \{5688\}$, der U er definert på side 9. Vi får $\overline{u^*} = \frac{1}{n}\sum_{i=1}^{8} u_i^* = 718$, og $\hat{M}_{U^*} = \frac{8+9}{2} = 8.5$. Dette illustrerer en viktig forskjell mellom gjennomsnitt og median. Det finnes ingen grense for hvor mye én enkelt observasjon kan påvirke gjennomsnittet. Legges det til en ekstra observasjon vil endringen i medianen kun påvirkes av en av naboobservasjonene til den eller de observasjoner som var i midten fra før. Uteliggeres størrelse har ingen innvirkning på den medianen, men de kan ha stor innvirkning på gjennomsnittet. Vi sier derfor at empirisk median generelt er et mer *robust* mål for sentralitet enn gjennomsnitt. Det samme gjelder som tidligere nevnt for kvantiler som for medianen, størrelsen på uteliggere spiller ikke inn.

Uteliggende observasjoner er i noen tilfeller sjeldne, og det kan være liten sjanse for at de inntreffer i et datasett. I noen situasjoner kan uteliggere være feilmålinger. I andre situasjoner, som for eksempel dersom vi har et datasett som viser de ansattes inntekter i en bedrift, kan vi ha en fordeling som er høyreskjev (for en definisjon av høyreskjevhet, se seksjon 2.5.4). Her kan enkelte observasjoner være mye større enn de fleste andre. I alle slike situasjoner vil vi kunne få en betydelig forskjell på de to sentralmålene, da medianen ikke vil ta hensyn til størrelsen blant eventuelle uteliggere, mens gjennomsnittet derimot vil la seg påvirke. Hvis vi har en symmetrisk fordeling vil forventningsverdien og den teoretiske medianen som tidligere nevnt være identiske, og da vil det i de fleste situasjoner være enklest å jobbe med gjennomsnittet i den empiriske situasjonen. Ved skjeve fordelinger vil medianen kunne være å foretrekke som sentralestimat.

I situasjoner hvor man har uteliggere kan man velge å rett og slett se bort fra dem når man skal beregne gjennomsnitt, ved for eksempel å se bort fra et visst antall av de største og eventuelt de minste observasjonene. Dette kan løse problemet ved at uteliggere får for stor innflytelse, men samtidig vil det være uheldig å se bort fra observasjoner, slik at å bruke median i stedet vil kunne sies å være en bedre løsning.

2.5 Spredning og skjevhet

I tillegg til sentrumsverdier er vi som regel interessert i hvor mye observasjoner varierer i størrelse. Det mest brukte målet for spredning er standardavvik, som bygger på teorien for forventning/gjennomsnitt. Dette målet lar seg påvirke av uteliggere. Et alternativt mål, som vi kan sette i sammenheng med teorien for median og kvantiler, er interkvartil rekkevidde, som ikke lar seg påvirke av uteliggere.

2.5.1 Standardavvik

Standardavviket er representert ved σ (σ^2 er det vi kaller varians). Empirisk standardavvik er representert ved henholdsvis s_n og $\hat{\sigma_n}$, se under.

Vi har en stokastisk variabel X. Vi lar μ være den sanne forventningsverdien $\mathrm{E}(X)$, og $f(x)$ sannsynlighetstettheten i det kontinuerlige tilfellet, og $P(X = x)$ punktsannsynligheten i det diskrete tilfellet[3].

$$\text{I det kontinuerlige tilfellet:} \quad \sigma = \sqrt{\mathrm{E}[(X-\mu)^2]} = \sqrt{\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx}.$$

$$\text{I det diskrete tilfellet:} \quad \sigma = \sqrt{\mathrm{E}[(X-\mu)^2]} = \sqrt{\sum_{\forall x}(x-\mu)^2 P(X=x)}.$$

$$\text{Empirisk:} \quad s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Vi ser av formlene at det dreier seg om kvadratroten av et gjennomsnitt av kvadratavstand. I neste kapittel skal vi se hvordan kvadratavstand og gjennomsnitt hører sammen, med andre ord er standardavvik et spredningsmål som bygger på teorien for forventning/gjennomsnitt.

Dersom vi deler med n i stedet for $n-1$ i formelen for empirisk standardavvik, så får vi en alternativ estimator som vi her kaller $\hat{\sigma}_n$, men den mest brukte er s_n. At vi helst vil dele på $n-1$ forklares ved at observasjonene i gjennomsnitt ligger nærmere estimatet $\overline{x}$ enn den sanne verdien μ, slik at om vi deler på n vil vi få en estimator som vi forventer er mindre enn det sanne standardavviket σ. Vi kan vise at s_n^2 er forventningsrett, se beviset øverst på neste side. Derimot er ikke $\sqrt{s_n^2} = s_n$ forventningsrett, vi husker fra seksjon 2.3.3 at vi generelt har $\mathrm{E}[h(X)] \neq h[\mathrm{E}(X)]$, og dette gjelder altså når h er kvadratrotfunksjonen, det vil si $\mathrm{E}[\sqrt{X})] \neq \sqrt{[\mathrm{E}(X)]}$. For mer om dette, se [Johnson m. fl 1994] kapittel 13, seksjon 8.2.

[3] Se [Walpole m.fl 2002] på side 96.

Bevis for at s_n^2 er forventningsrett:[4]

$$\begin{aligned}
\mathrm{E}(s_n^2) &= \mathrm{E}(\tfrac{1}{n-1}[\textstyle\sum_{i=1}^n (x_i - \overline{x})^2]) \\
&= \mathrm{E}(\tfrac{1}{n-1}[\textstyle\sum_{i=1}^n x_i^2 - n\overline{x}^2]) \\
&= \tfrac{1}{n-1}[n\mathrm{E}(X_1^2) - n\mathrm{E}(\overline{x}^2)] \\
&= \tfrac{1}{n-1}[n(\sigma^2 + \mu^2) - n(\tfrac{\sigma^2}{n} + \mu^2)] \\
&= \sigma^2.
\end{aligned}$$

Standardavvik er et godt mål for spredning for symmetriske fordelinger, men det kan være vanskelig å tolke for skjeve fordelinger.

2.5.2 Interkvartil rekkevidde (IQR)

Når vi regner med kvantiler har vi et annet mål for spredning som er mye brukt, nemlig interkvartil rekkevidde (IQR: Interquartile Range):

$$\mathrm{IQR} = Q^{(0.75)} - Q^{(0.25)}.$$

Empirisk får vi da notasjonen $\widehat{\mathrm{IQR}} = \hat{Q}^{(0.75)} - \hat{Q}^{(0.25)}$. Vi kan tilsvarende definere andre spredningsmål, og bruke dem sammen for å få et mer komplett bilde på spredning, for eksempel vil $Q^{(0.975)} - Q^{(0.025)}$ gi et bilde på i hvilken grad vi har store haler. Interkvartil rekkevidde alene er dog det mest brukte spredningsmålet av denne typen.

Det finnes enda flere mulige mål for spredning, for eksempel gjennomsnittlig absoluttavstand i stedet for kvadratroten av gjennomsnittlig kvadratavstand, som vi hadde for standardavvik. Vi vil i kapittel 3 se hvordan kvadratavstand henger sammen med gjennomsnitt, og hvordan absoluttavstand er relatert til median.

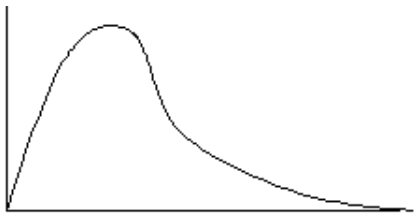
2.5.3 Sammenligning av standardavvik og IQR

Har vi normalfordeling, vil regionen avgrenset av sentrum og ett standardavvik til for eksempel til høyre for sentrum dekke 34.13 % av total sannsynlighetsmasse (dette kan leses ut av en normalfordelingstabell). For andre fordelinger gjelder ikke nødvendigvis dette. IQR dekker de 50 % av sannsynlighetsmassen som er i midten uansett fordeling, dvs 25 % av sannsynlighetsmassen skal ligge utenfor på hver sin side, slik at i situasjonen med normalfordeling vil IQR være omtrent halvannen ganger så stor som standardavviket. Har vi uteliggere vil IQR ikke bli påvirket av disse, men det er ingen begrensning på hvor stor effekt de kan ha på standardavviket.

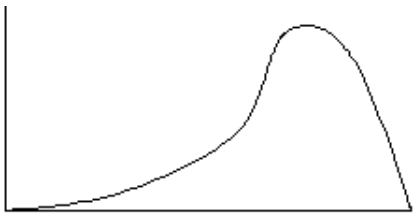
[4] Hentet fra [Casella og Berger 2002], side 214.

2.5.4 Skjevhet

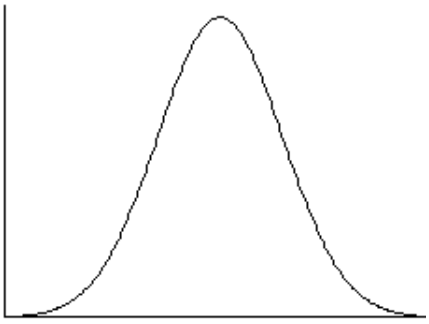
Standardavviket forteller oss om spredningen til en fordeling, men ikke om eventuell ikke-symmetri. Her kommer skjevhet inn i bildet, og vi regner ut skjevhet ved tredje sentralmoment[5] om forventningsverdien, $\int_{-\infty}^{\infty}(x-\mu)^3 f(x)dx$ i det kontinuerlige tilfellet, og $\sum_{\forall x}(x-\mu)^3 P(X=x)$ i det diskrete tilfellet. Normalt standardiseres dette ved å dele på σ^3. Empirisk blir det $\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{s_n^3}$, og er datapunktene symmetrisk plassert rundt gjennomsnittet $\overline{x}$, så vil det tredje standardiserte sentralmomentet være 0. Er skjevheten positiv, indikerer det at høyre hale er lengre enn venstre, og vi sier vi har høyreskjevhet (omgrepet positiv skjevhet brukes også om høyreskjevhet). Dette betyr at man for en stokastisk variabel med høyreskjev fordeling vil forvente å få en del observasjoner som er en god del større enn de fleste observasjonene. Dette ser vi på Figur 2.1 under, som viser eksempel på tettheten til en høyreskjev fordeling. For negativt tredje sentralmoment har vi tilsvarende venstreskjevhet (eller negativ skjevhet), som blir det motsatte.



Figur 2.1: *Høyreskjev fordeling.*



Figur 2.2: *Venstreskjev fordeling.*



Figur 2.3: *Symmetrisk fordeling.*

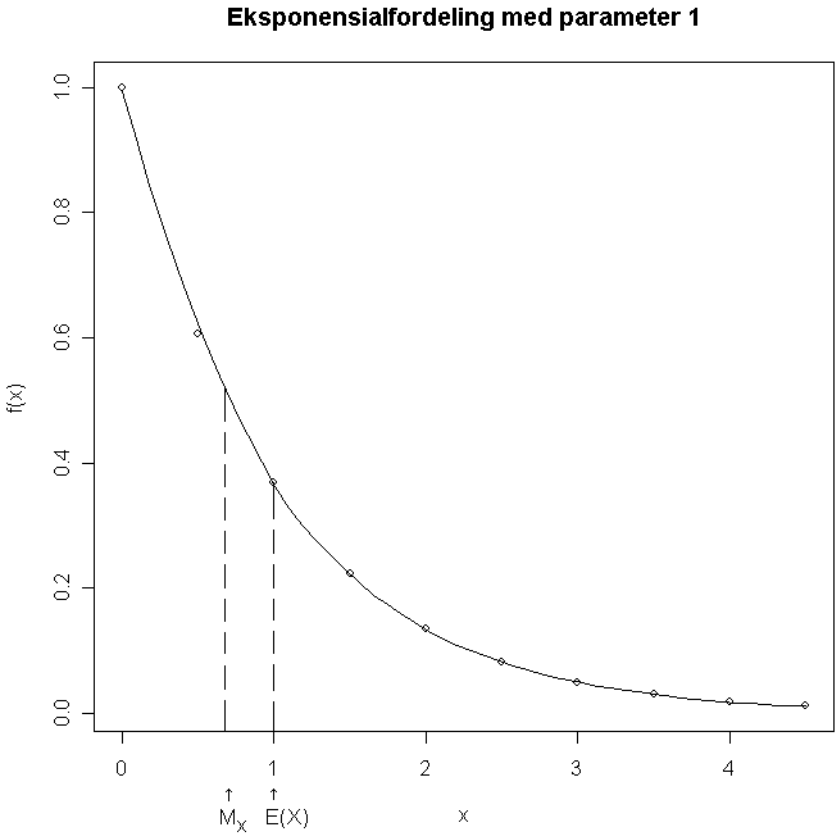
En alternativ, kvantilbasert metode eksisterer også for å definere skjevhet, se [Hao og Naiman 2007] side 12-14 for mer om dette.

[5]Forventning er førstemomentet og varians andre sentralmoment, se [Casella og Berger 2002] side 59-61.

2.6 Teoretisk eksempel: Eksponensialfordeling

Vi ser på en enkel eksponensialfordeling med parameter 1, dvs vi har tettheten:

$$f(x) = \left\{ \begin{array}{ll} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{array} \right\}.$$



Figur 2.4: *Eksponensialfordeling med tetthet $f(x) = e^{-x}$. Medianen og forventningsverdien er markert.*

Forventningsverdi:

$$\begin{aligned} \mathrm{E}(X) = \int_0^\infty x e^{-x} dx &= \tfrac{1}{2} x^2 e^{-x} |_0^\infty - \int_0^\infty -e^{-x} dx \\ &= -[e^{-x}]_0^\infty \\ &= 1. \end{aligned}$$

Median:

$$\begin{aligned} F(M_X) &= \tfrac{1}{2} \\ \int_0^{M_X} e^{-x}dx &= \tfrac{1}{2} \\ 1 - e^{-M_X} &= \tfrac{1}{2} \\ e^{-M_X} &= \tfrac{1}{2} \\ -M_X &= \ln \tfrac{1}{2} \\ M_X &= \ln 2 \approx 0.69. \end{aligned}$$

Vi ser at medianen er betydelig mindre enn forventningsverdien. Dette var ventet, for vi ser på Figur 2.4 på forrige side at vi har en høyreskjev fordeling, der den lange høyrehalen preger forventningsverdien, og sørger for at denne blir større enn medianverdien.

Vi kan regne ut hvilken kvantil forventningsverdien ligger på:

$$\begin{aligned} \tau_{\mathrm{E}(X)} &= F(\mathrm{E}(X)) \\ \tau_{\mathrm{E}(X)} &= F(1) \\ \tau_{\mathrm{E}(X)} &= 1 - e^{-1} \\ \tau_{\mathrm{E}(X)} &\approx 0.63. \end{aligned}$$

Det vil si at nesten omtrent 63 % av sannsynlighetsmassen er mindre enn forventningsverdien i denne situasjonen, dvs høyreskjevheten gjør at den forventede verdien er mye større enn verdiene som ligger rundt midten av fordelingen.

Vi kan også sammenligne standardavvik med interkvartil rekkevidde. Vi hadde $\mathrm{E}(X) = 1$, noe som gir oss følgende standardavvik[6]:

$$\sigma = \sqrt{\int_0^\infty (x-1)^2 e^{-x} dx} = 1.$$

Interkvartil rekkevidde[7]:

$$\mathrm{IQR} = Q^{(0.75)} - Q^{(0.25)} = \ln 4 - (\ln 4 - \ln 3) = \ln 3 \approx 1.10.$$

Vi ser at den skjeve fordelingen gjør at standardavviket blir større, relativt til interkvartil rekkevidde, sammenlignet med situasjonen for eksempel for en normalfordeling (se side 15), hvor vi hadde $\mathrm{IQR} \approx 1.5 \cdot \sigma$.

[6] Se [Casella og Berger 2002] side 59-60, utledning er tilsvarende som for $\mathrm{E}(X)$.

[7] Vi finner $Q^{(0.75)}$ og $Q^{(0.25)}$ på samme måte som vi fant medianen øverst på siden.

2.7 Empirisk eksempel: Skatteliste

Inntekt, skatt og formue er eksempler på noe som har ikke-symmetrisk fordeling, fordelingene er som regel mer eller mindre høyreskjeve. Dette gjør at gjennomsnittsverdien og medianen i et datasett kan være svært forskjellige fra hverandre av grunner vist for eksempel i seksjon 2.4 på side 13. Vi skal her utføre noen enkle beregninger på datasettet som er vedlagt i appendikset i seksjon A.1, nemlig en skatteliste bestående av skattbar inntekt, formue og skatt for 241 menn og 234 kvinner. Se tabellen under.

Menn	Gjennomsnitt	Median	Standardavvik	Interkvartil rekkevidde
Inntekt	269 183	210 977	269 762	257 025
Formue	396 537	96 653	657 191	570 985
Skatt	95 133	67 393	119 324	110 757
Kvinner	Gjennomsnitt	Median	Standardavvik	Interkvartil rekkevidde
Inntekt	130 953	105 229	95 812	84 181
Formue	238 633	55 908	653 360	335 521
Skatt	35 676	24 610	38 821	40 036

Tabell 2.1: *Empirisk gjennomsnitt og median for henholdsvis inntekt, skatt og formue, samt tilhørende empiriske variasjonsmål, alt rundet av til nærmeste krone.*

Vi observerer her blant annet følgende:

- Det er moderat forskjell på gjennomsnitt og median for inntekt og skatt. Gjennomsnittet er i begge tilfeller noe større, noe som tyder på en tanke høyreskjevhet.
- For formue blir det radikale forskjeller, noe som kommer av en del uteliggende observasjoner i høyre hale.
- Standardavvik er større enn interkvartil rekkevidde overalt, noe som er et tegn på at vi har skjeve fordelinger. Ved normalfordeling skal interkvartil rekkevidde være omtrent halvannen ganger så stor som standardavviket, se seksjon 2.5.3.

Vi skal bruke datasettet i A.1 til å vise eksempler på kvantilregresjon senere, se seksjon 4.6.

3 Gjennomsnitt, median og kvantiler som løsning av minimaliseringsproblem

Dette kapitlet er basert på [Hao og Naiman 2007], side 15-19.

Vi skal her vise hvordan vi kan sette opp minimaliseringsproblem som gir oss henholdsvis gjennomsnitt og empirisk median. Vi vil se at gjennomsnittet minimerer kvadratavstand, mens den empiriske medianen minimerer absoluttavstand. Metoden for den empiriske medianen generaliseres så til å gjelde alle kvantiler ved å innføre en type vektet absoluttavstand. Det er ved bruk av slike minimaliseringsproblem vi regner oss fram til estimatorer til regresjonskoeffisienter, noe vi skal se på i kapittel 4.

3.1 Gjennomsnitt

3.1.1 Empirisk

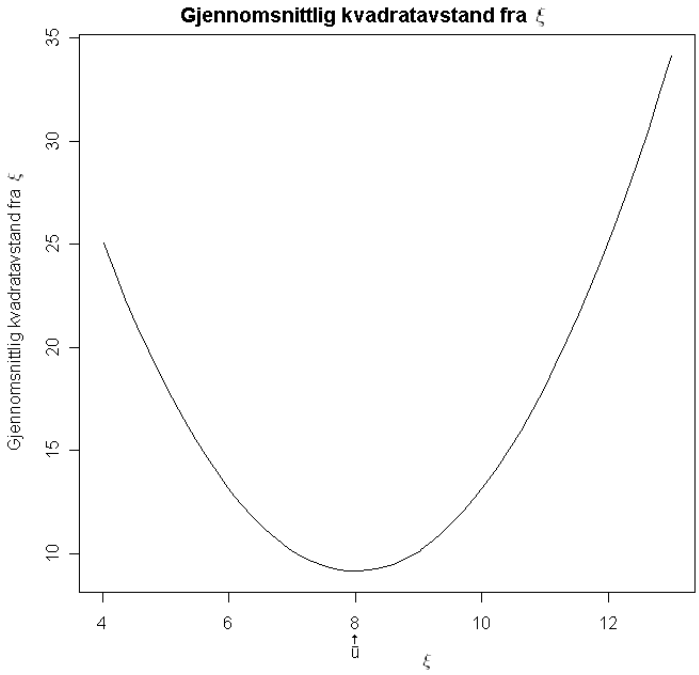
Gitt en mengde av observasjoner $X : \{x_1, x_2, ..., x_n\}$.

Vi har at $\xi = \overline{x}$ minimerer gjennomsnittlig kvadratavstand $\frac{1}{n}\sum_{i=1}^{n}(x_i - \xi)^2$.

Bevis:

$$\begin{aligned}
\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \xi)^2 &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}((x_i - \overline{x}) + (\overline{x} - \xi))^2 \\
&= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2 + \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(\overline{x} - \xi)^2 + 2(\overline{x} - \xi)\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \overline{x}) \\
&= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2 + \tfrac{1}{n}(\overline{x} - \xi)^2 n + 2(\overline{x} - \xi)\tfrac{1}{n} \cdot 0 \\
&= \hat{\sigma}_n^2 + (\overline{x} - \xi)^2, \text{ der } \hat{\sigma}_n^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \overline{x})^2.
\end{aligned}$$

Vi ser at uttrykket $\hat{\sigma}_n^2 + (\overline{x} - \xi)^2$ minimeres når $\xi = \overline{x}$, siden kvadrerte størrelser alltid er større eller lik 0 for reelle tall. Se også Figur 3.1 på neste side.



Figur 3.1: *Ser på mengden U fra seksjon 2.1.1 på side 9:* $f(\xi) = \frac{1}{7}\sum_{i=1}^{7}(u_i - \xi)^2$.

Vi ser at vi får bunnpunkt for $f(\xi)|_{\xi=\overline{u}}$.

Senere skal vi se at det er nettopp løsning av minimaliseringsproblem av denne typen som bestemmer estimatorene til regresjonskoeffisientene i ordinær regresjonsanalyse, noe som er beskrevet i seksjon 4.2. Ordinær regresjon er altså gjennomsnittbasert.

3.1.2 Teoretisk

Vi har at $\xi = \mathrm{E}[X]$ minimerer forventet kvadratavstand $\mathrm{E}[(X - \xi)^2]$.

Bevis:

$$\begin{aligned}
\mathrm{E}[(X-\xi)^2] &= \mathrm{E}[X^2 - 2\xi X + \xi^2] \\
&= \mathrm{E}[X^2] - 2\xi\mathrm{E}[X] + \xi^2 + ((\mathrm{E}[X])^2 - (\mathrm{E}[X])^2) \\
&= (\xi^2 - 2\xi\mathrm{E}[X] + (\mathrm{E}[X])^2) + (\mathrm{E}[X^2] - (\mathrm{E}[X])^2) \\
&= (\xi - \mathrm{E}[X])^2 + \mathrm{Var}(X).
\end{aligned}$$

Siden $\mathrm{Var}(X)$ er konstant i ξ, og $(\xi - \mathrm{E}[X])^2 \geq 0$, så har vi at forventet kvadratavstand blir minimert når vi setter $\xi = \mathrm{E}[X]$.

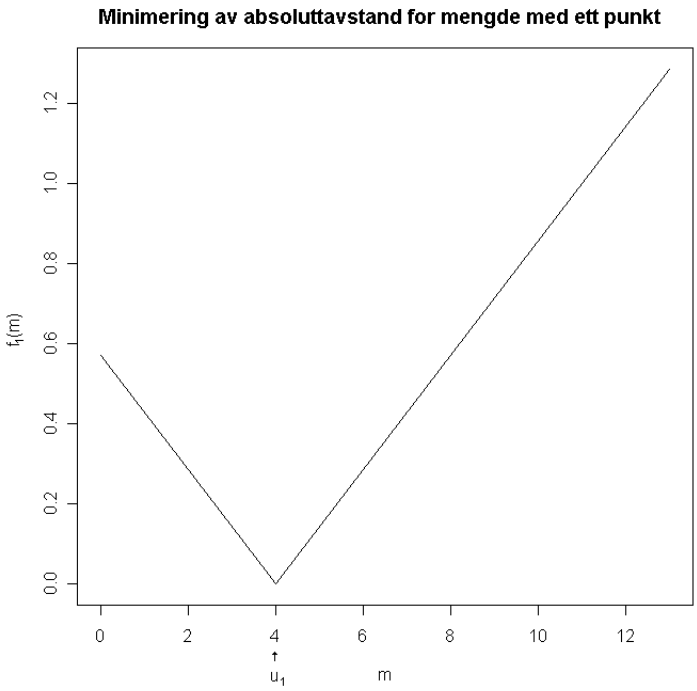
3.2 Median

3.2.1 Empirisk

Vi forholder oss til en endelig mengde $X : \{x_1, x_2, ..., x_n\}$. Vi har at $m = \hat{M}_X$ minimerer gjennomsnittlig absoluttavstand $\frac{1}{n}\sum_{i=1}^{n}|x_i - m|$.

Begrunnelse:

Vi ser at $\frac{1}{n}\sum_{i=1}^{n}|x_i - m|$ er en sum av n funksjoner $f_i(m) = \frac{|x_i - m|}{n}$, og hver av disse funksjonene er V-formet. Vi har et knekkpunkt i $(x_i, 0)$, og en rett linje som stiger i begge retninger fra knekkpunktet, se figur 3.2 under.



Figur 3.2: *Ser på mengden U fra seksjon 2.1.1:* $f_1(m) = \frac{|u_1 - m|}{n}$, *der* $u_1 = 4$ *og* $n = 7$.

Vi ser at $f_1(m)$ minimeres til verdien 0 for $m = u_1$. Siden bunnpunktet er et knekkpunkt finnes det ikke noen derivert her, men så lenge vi har en konveks[8] funksjon vi vil vi ha bunnpunkt enten når den deriverte er 0 eller når de to retningsderiverte har ulikt fortegn.

[8]Med en *konveks* funksjon mener vi en funksjon som har den egenskapen at om vi trekker en rett linje mellom to vilkårlig valgte punkt på kurven til funksjonen, så vil kurven alltid ligge under (eller rett på) denne linjen mellom disse to punktene. En slik funksjon vil derfor ha én bunnverdi, som er global, men som vi senere skal se kan det være mange punkter som har denne bunnverdien.

Vi har generelt $\frac{d}{dx}|x| = \frac{x}{|x|}$, som gir $\frac{d}{dm}f_i(m) = \frac{d}{dm}\frac{|x_i-m|}{n} = \frac{1}{n}\frac{(x_i-m)}{|x_i-m|}(-1)$ med kjerneregelen, og vi ser at dette gir at $\frac{d}{dm}f_i(m)$ er $-\frac{1}{n}$ for $m < x_i$ og $\frac{1}{n}$ for $m > x_i$. Det finnes ingen derivert i punktet $m = x_i$, men vi har retningsderiverte, nemlig $-\frac{1}{n}$ i negativ retning og $\frac{1}{n}$ i positiv retning, som forteller oss at $f_i(x_i)$ er et bunnpunkt.

Funksjonen $f(m) = \frac{1}{n}\sum_{i=1}^{n}|x_i-m|$ er summen av alle de n delfunksjonene $f_i(m)$, og det kan vises at vi vil få en stykkevis lineær, konveks funksjon med knekkpunkt alle steder hvor hver enkelt delfunksjon har knekkpunkt, med andre ord knekkpunkt for $m = x_i$ for $i = 1, 2, ..., n$ (når vi videre snakker om i mener vi for alle i, som her).

Vi definerer v og h til å være antall observasjoner henholdsvis til venstre og til høyre for m. I [Hao og Naiman 2007][9] på side 17 skriver de, uten å begrunne nevneverdig, at de retningsderiverte til $f(m)$ i $m = x_i$ er $\frac{(v-h)}{n}$ i negativ retning og $\frac{(h-v)}{n}$ i positiv retning, og at dette gjør at f blir minimert når m har like mange punkt til høyre som til venstre for seg, altså når $m = \hat{M}_X$. Dette må være feil. Når vi har $v = h$, som vi har for den empiriske medianen, så får vi riktig nok at den retningsderiverte blir 0 i begge retninger i følge disse formlene, noe som gir bunnpunkt, men bare hvis vi har jevnt antall observasjoner (mer om dette under). Men for alle verdier av m som ikke har like mange punkter på hver side, så vil vi få forskjellig fortegn for de to retningsderiverte i følge formlene over, noe som også signaliserer bunnpunkt for en konveks funksjon. Dette betyr at i følge disse formlene for retningsderiverte så får vi bunnpunkt overalt, så disse formlene for retningsderiverte kan ikke stemme.

Jeg har kommet fram til at de retningsderiverte i stedet må bli $\frac{v-(h+1)}{n}$ i negativ retning og $\frac{(v+1)-h}{n}$ i positiv retning, for $m = x_i$. Dette stemmer for de retningsderiverte til $f_i(m)$ i punktet $m = x_i$, som var $-\frac{1}{n}$ i negativ retning og $-\frac{1}{n}$ i positiv retning. For $m \neq x_i$ blir de retningsderiverte i begge retninger $\frac{v-h}{n}$. De retningsderiverte til $f_i(m)$ i alle punkt $m < x_i$ er, som vi ser av Figur 3.2 på forrige side, $-\frac{1}{n}$ i begge retninger, og $\frac{1}{n}$ i begge retninger for $m > x_i$. Alt dette stemmer med formlene. Vi vil på de neste sidene foreta en detaljert drøfting for å begrunne formlene for $f(m)$.

Hvis vi tar den retningsderiverte for eksempel til venstre (dvs i negativ retning) i et punkt x_i, så vil x_i være til høyre for den retningsderiverte, og det er det som er grunnen til at vi skriver $h + 1$ for antall punkt som er til høyre for den retningsderiverte i retning venstre i et punkt, og tilsvarende $v+1$ den andre veien. $v-(h+1)$ blir da antall punkt til venstre for x_i minus antall punkt til høyre for x_i, men så må vi trekke fra 1 ekstra, siden x_i selv vil ligge til høyre for den retningsderiverte i $m = x_i$. For $m \neq x_i$ kommer vi ikke borti noe sånt problem, her blir det bare $\frac{v-h}{n}$ som retningsderivert i begge retninger, altså fremdeles antall punkt til venstre for den retningsderiverte minus antall punkt til høyre. Dette blir en generalisering av

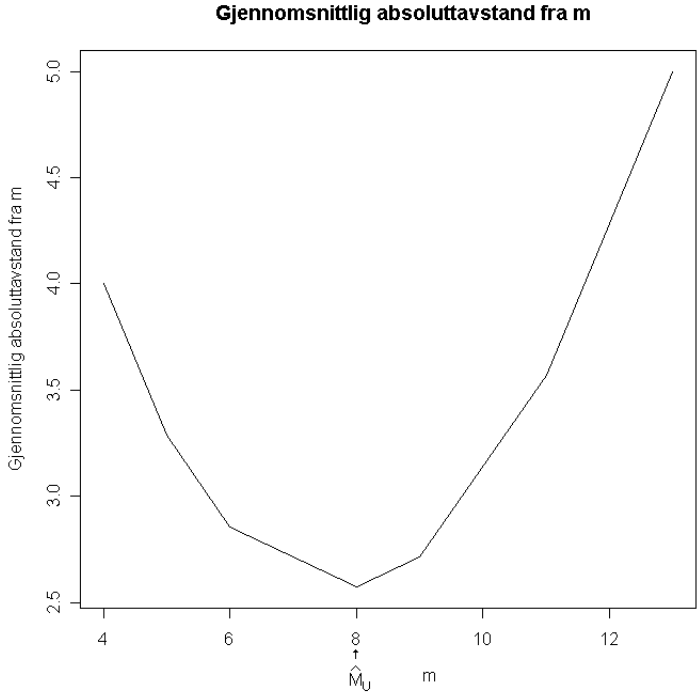
[9] De bruker litt annen notasjon, bl.a. s og r i stedet for henholdsvis h og v som er brukt her.

tilfellet for bare ett punkt, der vi har $v = h = 0$, og da får vi at den retningsderiverte i negativ retning er $-\frac{1}{n}$, og i positiv retning $\frac{1}{n}$. Vi har da bare ett punkt, og beveger vi oss til venstre er punktet på høyre side, og beveger vi oss til høyre er punktet på venstre side, slik at dette stemmer med formlene i avsnittet over.

Vi kan se ut fra dette at $f(m)$ er en konveks funksjon, som har bunnpunkt i $f(m)|_{m=\hat{M}_X}$, fordi vi i punktet $m = \hat{M}_X$ vil ha $v = h$ siden den empiriske medianen er definert til å ha like mange punkt til venstre for seg som til høyre for seg. Når vi skal vise dette må det skilles mellom når vi har henholdsvis odde og jevnt antall observasjoner. Den empiriske medianen vil gi bunnpunkt i begge tilfeller, men det vil skje på to ulike måter, se de to eksemplene under.

Odde antall observasjoner

i	1	2	3	4	5	6	7
u_i	4	5	6	**8**	9	11	13
$f(m)\|_{m=u_i}$	4	23/7	20/7	**18/7**	19/7	25/7	35/7
Retningsderivert mot venstre	-1	-5/7	-3/7	**-1/7**	1/7	3/7	5/7
Retningsderivert mot høyre	-5/7	-3/7	-1/7	**1/7**	3/7	5/7	1



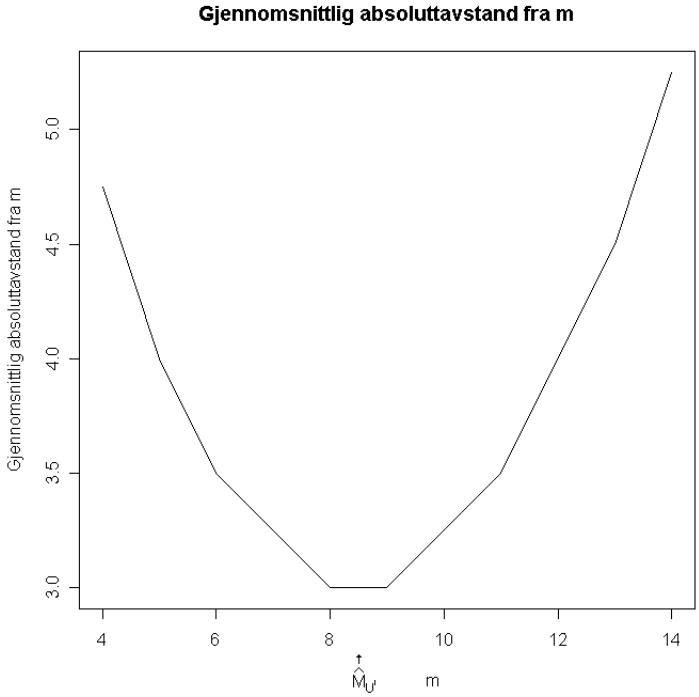
Figur og tabell 3.3: *Ser på mengden U fra seksjon 2.1.1:* $f(m) = \frac{1}{7}\sum_{i=1}^{7}|u_i - m|$.

Når vi har odde antall observasjoner, så vil de retningsderiverte i det empiriske medianpunktet være henholdsvis $-\frac{1}{n}$ i negativ retning og $\frac{1}{n}$ i positiv retning. Når de retningsderiverte i et punkt har forskjellig fortegn, og funksjonen er konveks, så må vi ha et bunnpunkt. At funksjonen er konveks kommer av at de retningsderiverte øker i absoluttverdi jo lenger fra bunnpunktet på hver side de befinner seg.

Se Figur og tabell 3.3 på forrige side for et eksempel hvor vi har odde antall observasjoner. Vi husker fra seksjon 2.2 at vi for U har $\hat{M}_U = u_{\frac{n+1}{2}} = u_4 = 8$. Vi ser både fra tabellen og grafen at $f(m)$ har bunnpunkt i $f(m)|_{m=\hat{M}_U}$. Vi ser ellers at formlene for retningsderiverte fra side 23 i de forskjellige punktene stemmer med det vi ser i tabellen og på figuren.

Jevnt antall observasjoner

i	1	2	3	4	5	6	7	8
u_i	4	5	6	**8**	**9**	11	13	14
$f(m)\|_{m=u_i}$	4.75	4	3.5	**3**	**3**	3.5	4.5	5.25
Retningsderivert mot venstre	-1	-6/8	-4/8	**-2/8**	**0**	2/8	4/8	6/8
Retningsderivert mot høyre	-6/8	-4/8	-2/8	**0**	**2/8**	4/8	6/8	1



Figur og tabell 3.4: *Ser på mengden $U' = U \bigcup \{14\}$, der U er definert i seksjon 2.1.1: Vi får $M_{U'} = 8.5$ og $f(m) = \frac{1}{8}\sum_{i=1}^{8} |u_i' - m|$.*

Når vi har jevnt antall observasjoner vil vi i følge formlene vi kom fram til på side 23 få retningsderiverte $-\frac{2}{n}$ i negativ retning og 0 i positiv retning for punktet $x_{\frac{n}{2}}$, samt 0 i negativ retning og $\frac{2}{n}$ i positiv regning for punktet $x_{(\frac{n}{2}+1)}$. I alle punkt mellom $x = x_{\frac{n}{2}}$ og $x = x_{(\frac{n}{2}+1)}$ vil vi i følge formelen $\frac{v-h}{n}$ få retningsderiverte lik 0, vi vil da ha en derivert lik 0, og i og med at funksjonen er konveks har vi da et bunnpunkt. Det følger at vi får en *bunnlinje* fra $x = x_{\frac{n}{2}}$ til $x = x_{(\frac{n}{2}+1)}$, og den empiriske medianen, som er definert til å være midtpunktet på denne linjen, vil altså ha bunnverdien.

Se Figur og tabell 3.4 på forrige side for et eksempel hvor vi har jevnt antall observasjoner, nemlig 8. Vi ser at hele linjen fra observasjon $f(u_4^{'})$ til $f(u_5^{'})$ har bunnverdien, og at den empiriske medianen er midtpunktet på denne linjen. De retningsderiverte i tabellen stemmer ellers med formlene, og vi ser også der at vi får bunnverdi for $u_4^{'}$ og $u_5^{'}$. De retningsderiverte til henholdsvis høyre og venstre er nemlig 0, noe som igjen indikerer en bunnlinje mellom disse to observasjonene, hvis midtpunkt er den empiriske medianen. Minimaliseringsproblemet gir oss altså her flere bunnpunkter, og den empiriske medianen, som er definert til å være det midterste punktet på linjen, gir oss et av disse[10].

Vi vil senere se at det er minimaliseringsproblem av denne typen som bestemmer estimatorene til regresjonskoeffisientene i medianregresjon (se seksjon 4.3). Utvides dette til å gjelde alle kvantiler ender vi opp med kvantilregresjon. Vi vil se at det å regne seg fram til disse estimatorene er mer krevende enn tilsvarende for ordinær regresjon.

3.2.2 Teoretisk

Siden begrunnelsen i det empiriske tilfellet er av noe intuitiv karakter, tar vi med et bevis (se [Hao og Naiman 2007] side 21-22) for det teoretiske tilfellet.

Vi har at $m = M_X$, medianen, minimerer forventet absoluttavstand $\mathrm{E}|X - m|$.

Bevis:
Vi ser på en kontinuerlig sannsynlighetsfordeling med deriverbar kumulativ fordelingsfunksjon F. Vi vil her kun gi et bevis for det kontinuerlige tilfellet.

$$\begin{aligned} \mathrm{E}|X-m| &= \int_{-\infty}^{\infty} |x-m| f(x)dx \\ &= \int_{-\infty}^{m} (m-x)f(x)dx + \int_{m}^{\infty} (x-m)f(x)dx. \end{aligned}$$

[10] Alle punktene på linjen vil telle som en løsning, noe som vil videreføres i kvantilregresjon ved at man regner alle løsninger som likeverdige. Man bruker ikke energi på å finne noe midtpunkt i mengden løsninger når vi ikke har unik løsning, men godtar den første løsningen man finner.

Fra det empiriske tilfellet har vi at dette er en konveks funksjon. Vi har altså kun én ekstremverdi, en bunnverdi. Vi partiellderiverer med hensyn på m og setter den partiellderiverte lik 0 for å finne en verdi av m som løser minimaliseringsproblemet.

Den partiellderiverte av det første leddet er:

$$\begin{aligned}\tfrac{\partial}{\partial m}\int_{-\infty}^{m}(m-x)f(x)dx &= \int_{-\infty}^{m}\tfrac{\partial}{\partial m}(m-x)f(x)dx + (m-x)f(x)|_{x=m} \\ &= \int_{-\infty}^{m} f(x)dx = F(m).\end{aligned}$$

Den partiellderiverte av det andre leddet er:

$$\begin{aligned}\tfrac{\partial}{\partial m}\int_{m}^{\infty}(x-m)f(x)dx &= \int_{m}^{\infty}\tfrac{\partial}{\partial m}(x-m)f(x)dx - (x-m)f(x)|_{x=m} \\ &= -\int_{m}^{\infty} f(x)dx = -(1-F(m)).\end{aligned}$$

Kombinerer vi disse to uttrykkene får vi:

$$\tfrac{\partial}{\partial m}\mathrm{E}|X-m| = F(m) - (1-F(m)) = 2F(m) - 1.$$

Settes dette lik 0 får vi $F(m) = 0.50$, noe som forteller oss at $m = M_X$, den teoretiske medianen, er en løsning av minimaliseringsproblemet.

3.3 Kvantiler

Vi kan generalisere situasjonen for medianen i forrige seksjon ved å innføre en form for vektet absoluttavstand (hentet fra [Hao og Naiman 2007] side 17-19) :

$$d_\tau(X,q) = \left\{ \begin{array}{ll} (1-\tau)|X-q|, & X < q \\ \tau|X-q|, & X \geq q \end{array} \right\}.$$

Vi er ute etter verdien av q som minimerer forventet vektet absoluttavstand, $\mathrm{E}[d_\tau(X,q)]$, og det viser seg at denne minimeres når $q = Q_X^{(\tau)}$.

Tilsvarende empirisk har vi at $q = \hat{Q_X}^{(\tau)}$ minimerer gjennomsnittlig vektet absoluttavstand[11]:

$$\tfrac{1}{n}\sum_{i=1}^{n} d_\tau(x_i,q) = \tfrac{1-\tau}{n}\sum_{x_i<q}|x_i-q| + \tfrac{\tau}{n}\sum_{x_i>q}|x_i-q|.$$

[11]Eventuelle punkter $x_i = q$ gir intet bidrag siden vi da har $|x_i - q| = 0$. Det samme prinsippet gjelder for tilsvarende formler for vektet absoluttavstand senere, henholdsvis på side 44 og 72.

Vi gir her kun et bevis for det teoretiske tilfellet, som er en generalisering av tilsvarende bevis for medianen.

Teoretisk bevis:

$E[d_\tau(X,q)] = (1-\tau)\int_{-\infty}^{q}(q-x)f(x)dx + \tau\int_{q}^{\infty}(x-q)f(x)dx.$

Som i situasjonen med medianen kan det vises at dette er en konveks funksjon. Det betyr at vi har kun én ekstremverdi, nemlig en bunnverdi. Vi partiellderiverer med hensyn på q og setter den partiellderiverte lik 0 for å finne en verdi av q som løser minimaliseringsproblemet.

Den partiellderiverte av det første leddet er:

$$\begin{aligned}(1-\tau)\tfrac{\partial}{\partial q}\int_{-\infty}^{q}(q-x)f(x)dx &= (1-\tau)[\int_{-\infty}^{q}\tfrac{\partial}{\partial q}(q-x)f(x)dx + (q-x)f(x)|_{x=q}] \\ &= (1-\tau)\int_{-\infty}^{q}f(x)dx = (1-\tau)F(q).\end{aligned}$$

Den partiellderiverte av det andre leddet er:

$$\begin{aligned}\tau\tfrac{\partial}{\partial q}\int_{q}^{\infty}(x-q)f(x)dx &= \tau[\int_{q}^{\infty}\tfrac{\partial}{\partial q}(x-q)f(x)dx - (x-q)f(x)|_{x=q}] \\ &= -\tau\int_{q}^{\infty}f(x)dx = -\tau(1-F(q)).\end{aligned}$$

Kombinerer vi disse to uttrykkene får vi:

$\frac{\partial}{\partial q}\mathrm{E}[d_\tau(X,q)] = (1-\tau)F(q) - \tau(1-F(q)) = F(q) - \tau.$

Settes dette lik 0 får vi $F(q) = \tau$, noe som forteller oss at $q = Q_X^{(\tau)}$, τ-kvantilen, er en løsning av minimaliseringsproblemet. Vi ser at om vi setter $\tau = 0.50$, får vi resultatet vi fikk for medianen i forrige seksjon, noe som viser at minimaliseringsproblemet med absoluttavstand er et spesialtilfelle av den generaliserte varianten vi har definert med vektet absoluttavstand for kvantiler. Minimaliseringsproblem av denne typen gir oss estimatorene til regresjonskoeffisientene i kvantilregresjon (se seksjon 4.4).

4 Regresjonsanalyse

Stoffet i dette kapitlet er hentet fra [Walpole m.fl 2002], [Hao og Naiman 2007] og [Koenker 2005].

Vi vil her forklare hva regresjonsanalyse er, samt vise hvordan den klassiske metoden med minste kvadraters estimater fungerer, og bruke dette som referansepunkt når vi går videre og ser på median- og kvantilregresjon.

Målsetningen i regresjonsanalyse er å modellere sammenhengen mellom en responsvariabel og forklaringsvariable. Ser vi på eksempelet i seksjon 2.7 på side 19, kan vi for eksempel være interessert i sammenhengen mellom formue og inntekt. Hva blir formuen (responsvariabel) når inntekt (forklaringsvariabel) er gitt? I en lineær regresjonsmodell vil vi, basert på de datapunktene som danner grunnlaget for modellen, få en regresjonslinje som viser estimert verdi av skatten for en gitt verdi av inntekt. Men hva slags estimat?

Vanlig regresjonsanalyse, som har vært i bruk en god stund, gir oss betingede gjennomsnittsestimater av forventningen til responsvariabelen ved å minimere kvadratavstand. Vi skal se hvordan dette fungerer, og videre reformulere og i stedet se på hvordan vi kan få betingede medianestimater av responsvariabelen, og vise at dette har en naturlig utvidelse i betingede kvantilestimater. Fordelen med den klassiske metoden med betinget gjennomsnitt er at det fungerer svært godt under ideelle forutsetninger, når responsvariabelen er normalfordelt eller tilnærmet normalfordelt, og når vi har konstant varians for alle verdier av en forklaringsvariabel.

Den alternative metoden med betinget median basert på å minimere absoluttavstand er vanskeligere å implementere. Det var ikke før på slutten av 70-tallet at man, grunnet bruk av datamaskiner, greide å gjennomføre dette i praksis. Fordelen med dette alternativet er at når vi har ikke-symmetrisk responsfordeling eller når forutsetningen om konstant varians feiler, så vil den kunne gi mer fornuftige estimater for sentralitet. Ikke minst gir den generelle metoden med betingede kvantiler oss muligheten til å analysere utviklingen ethvert sted i fordelingen.

Vi skal se på alt dette, og for å gjøre det enklest mulig å sette seg inn i ting skal vi konsentrere oss om en enkel, lineær regresjonsmodell. Metodene vi beskriver kan så generaliseres til mer avanserte modeller.

4.1 En enkel, lineær regresjonsmodell

Vi antar vi har en mengde datapunkter $[(x_i, y_i); i = 1, 2, ..., n]$. Ut fra denne mengden kan vi lage en enkel, lineær regresjonsmodell. En responsvariabel Y gitt x, der x er forklaringsvariabelen, fungerer som en stokastisk variabel:

$$Y|x = \beta_0 + \beta_1 x + \varepsilon.$$

Vi har i modellen β_0 og β_1, som er ukjente parametere som vi kaller regresjons-

koeffisientene, i dette tilfellet henholdsvis skjæringspunkt ved y-aksen og stigningstall. $Y|x$ er en betinget funksjon, som viser oss hvor mye Y vil endre seg når x endrer seg. Støyleddet ε står for den stokastiske delen av $Y|x$, og for ordinær minste kvadraters regresjonsanalyse antar vi støyleddet har normalfordeling med $\mathrm{E}(\varepsilon) = 0$ og $\mathrm{Var}(\varepsilon) = \sigma^2$, og de n støyleddene vi har er antatt å være uavhengige av hverandre, slik at uavhengighet også er antatt for responsvariablene[12]. For en gitt x-verdi får vi da at y-verdien, siden $\mathrm{E}(\varepsilon) = 0$, er normalfordelt rundt den "sanne" regresjonslinjen $\mathrm{E}(Y|x) = \beta_0 + \beta_1 x$, med $\mathrm{Var}(Y|x) = \sigma^2$. Dette er en antakelse som noen ganger svikter[13], både fordi vi kan ha skjev fordeling, og siden variansen kan variere etter som vi forandrer verdien til forklaringsvariabelen. I kvantilregresjon vil vi ikke anta noen spesiell fordeling for støyleddene, noe som er unikt i sammenheng med regresjonsanalyse. Man har andre former for regresjonsanalyse hvor man legger til rette for skjeve fordelinger, men man går ikke der bort fra forutsetningen om identisk, uavhengig fordelte støyledd (men altså ikke nødvendigvis normalfordelte), slik vi vil se at man gjør i kvantilregresjon. I de situasjonene kan det fremdeles oppstå problemer, dette hvis fordelingene ikke er som antatt, og som før hvis variansen ikke er konstant.

Merk at vi aldri kan finne ut hva den sanne regresjonslinjen eller variansen er, men vi kan bruke datapunktene til å estimere β_0 og β_1 (og også σ^2 når vi antar støyleddene er identisk, uavhengig fordelte) for slik å danne en estimert regresjonslinje, som vi i tilfellet minste kvadraters regresjon skriver som:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Den estimerte responsverdien for en gitt x-verdi er da $\hat{y}$. Det er måten vi finner estimatorene til β_0 og β_1 på som utgjør forskjellen på ordinær regresjon og kvantilregresjon. I sistnevnte tilfelle brukes notasjonen $\hat{\beta}_0^{(\tau)}$ og $\hat{\beta}_1^{(\tau)}$ for estimatorene, mens i spesialtilfellet medianen kan vi skrive $\tilde{\beta}_0$ og $\tilde{\beta}_1$, som altså tilsvarer $\hat{\beta}_0^{(0.50)}$ og $\hat{\beta}_1^{(0.50)}$.

4.2 Ordinær regresjon ved minste kvadraters metode

Vi husker fra seksjon 3.1 at gjennomsnitt kan skrives som et minimaliseringsproblem hvor man minimerer gjennomsnittlig kvadratavstand. Det samme prinsippet med minste kvadraters metode brukes i ordinær regresjon (her sier vi at vi minimerer total kvadratavstand, dvs vi lar være å dele på n, noe som ikke får noen betydning i praksis når vi skal finne hvilke verdier av parametrene som minimerer uttrykket),

[12] Vi har responsvariablene $\{Y_i\}, i = 1, 2, ..., n$, men vi forholder oss i regresjonsmodellen her og i modeller videre til én generell responsvariabel Y.

[13] Men holder denne antakelsen gir ordinær regresjon bedre resultat enn medianregresjon, dog er forskjellen da liten.

og vi sier at ordinær regresjon er gjennomsnittbasert.

Vi har de observerte responsverdiene $y_i, (i = 1, 2, ..., n)$, med tilhørende verdier av forklaringsvariabelen, og vi ønsker å finne estimater $\hat{y}_i$ som er best mulig, ut fra den enkle, lineære regresjonsmodellen som vist på forrige side. Den mest populære metoden er da å finne minste kvadraters estimatorer for β_0 og β_1. Den totale kvadratavstanden i y-retning mellom den estimerte regresjonslinjen og hvert datapunkt, SSE (sum of squared errors), skriver vi som:

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Vi ønsker å minimere SSE for $\hat{\beta}_0$ og $\hat{\beta}_1$, og vi ender da opp med estimatorene:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{og} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Bevis:

Funksjoner bestående av ledd på formen x^2 er konvekse med ett unikt ekstrempunkt, nemlig et bunnpunkt. Vi partiellderiverer med hensyn på henholdsvis $\hat{\beta}_0$ og $\hat{\beta}_1$ og setter lik 0 og løser ut, førstnevnte først:

$$\begin{aligned}
\tfrac{\partial}{\partial \hat{\beta}_0}\text{SSE} &= -2\textstyle\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
0 &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_i - \hat{\beta}_0 - \hat{\beta}_1 \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i \\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x}.
\end{aligned}$$

$$\begin{aligned}
\tfrac{\partial}{\partial \hat{\beta}_1}\text{SSE} &= -2\textstyle\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \\
0 &= \textstyle\sum_{i=1}^{n} x_i y_i - (\overline{y} - \hat{\beta}_1 \overline{x})\sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
0 &= -\hat{\beta}_1(\textstyle\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2) + \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) + \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \overline{x}\cdot\overline{y}}{\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n}(x_i - \overline{x}) + \sum_{i=1}^{n} \overline{x}^2} \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.
\end{aligned}$$

Dette er den enkleste formen for ordinær regresjon. Vi kan utvide modellen til for eksempel en multippel regresjonsmodell som inneholder flere forklaringsvariable, en modell som inneholder ikke-lineære ledd eller interaktiv effekt mellom to forklaringsvariable, eller en generalisert lineær modell, se eksempler under:

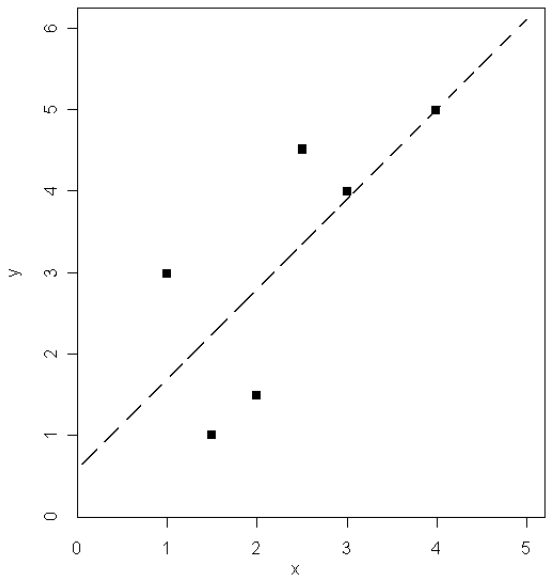
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon.$

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^2 + \beta_4 \log x_4 + \beta_5 x_1 x_2 + \varepsilon.$

$g(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$, der g er en monoton, deriverbar funksjon.

I alle disse tilfellene bruker man vanligvis samme prinsipp om minste kvadraters metode for å finne estimatorer til regresjonskoeffisientene. For detaljer om ordinær regresjonsanalyse ved minste kvadraters metode i disse tilfellene, se for eksempel [Walpole m.fl 2002] og [Dobson 2002].

Eksempel 4.1
Vi tenker oss at vi har et datasett med 6 observasjoner,
$(x_i, y_i) = [(1,3); (1.5,1); (2,1.5); (2.5,4.5); (3,4); (4,5)]$, der Y er responsvariabelen og X er forklaringsvariabelen. Vi får $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = 1.1$ og $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 0.6$. Figur 4.1 under viser regresjonslinjen og alle datapunktene.



Figur 4.1: *Datapunktene fra Eksempel 4.1 med minste kvadraters regresjonslinje.*

Vi skal senere i dette kapitlet bruke dette datasettet til å illustrere median- og kvantilregresjon, se Eksempel 4.2 på side 37 og Eksempel 4.4 på side 45.

4.2.1 Forventning og varians til estimatorene til regresjonskoeffisientene

Estimatorene er forventningsrette.

Bevis:

Vi har $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^n (x_i-\overline{x})^2} = \frac{\sum_{i=1}^n (x_i-\overline{x})Y_i}{\sum_{i=1}^n (x_i-\overline{x})^2}$, siden $\overline{y}\sum_{i=1}^n (x_i-\overline{x}) = 0$. Vi regner x-verdiene som gitt, slik at de blir behandlet som konstanter. Vi får:

$$\mathrm{E}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i-\overline{x})\mathrm{E}(Y_i|x_i)}{\sum_{i=1}^n (x_i-\overline{x})^2} = \frac{\sum_{i=1}^n (x_i-\overline{x})(\beta_0+\beta_1 x_i)}{\sum_{i=1}^n (x_i-\overline{x})^2} = \frac{\sum_{i=1}^n (x_i-\overline{x})(\overline{y}-\beta_1\overline{x}+\beta_1 x_i)}{\sum_{i=1}^n (x_i-\overline{x})^2} = \beta_1 \frac{\sum_{i=1}^n (x_i-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2} = \beta_1.$$

$$\mathrm{E}(\hat{\beta}_0) = \mathrm{E}(\overline{y} - \hat{\beta}_1\overline{x}) = y - \beta_1 x = \beta_0.$$

Vi regner så ut variansen til estimatorene:

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\frac{\sum_{i=1}^n (x_i-\overline{x})Y_i}{\sum_{i=1}^n (x_i-\overline{x})^2}) = \frac{\sum_{i=1}^n \mathrm{Var}((x_i-\overline{x})Y_i)}{(\sum_{i=1}^n (x_i-\overline{x})^2)^2} = \frac{\sum_{i=1}^n (x_i-\overline{x})^2\mathrm{Var}(Y_i)}{(\sum_{i=1}^n (x_i-\overline{x})^2)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}.$$

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\overline{y} - \hat{\beta}_1\overline{x}) = \frac{\sigma^2}{n} + \overline{x}^2\frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2} = \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i-\overline{x})^2}) = \sigma^2\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i-\overline{x})^2}.$$

Vi kan også estimere σ^2, som er variansen til støyleddene, og som utgjør en del av uttrykkene for varians over. Ved å bruke de elementære resultatene vi til nå har utledet kan vi vise at vi har den forventningsrette estimatoren:

$s^2 = \frac{\mathrm{SSE}}{n-2}$.

Bevis:

$$\begin{aligned}
\mathrm{E}(\mathrm{SSE}) &= \mathrm{E}[\textstyle\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2] \\
&= \mathrm{E}[\textstyle\sum_{i=1}^n (Y_i - \overline{Y} - \hat{\beta}_1(x_i - \overline{x}))] \\
&= \mathrm{E}[\textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (Y_i - \overline{Y})(x_i - \overline{x}) + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \overline{x})^2] \\
&= \mathrm{E}[\textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2 - 2\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \overline{x})^2 + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \overline{x})^2] \\
&= \mathrm{E}[\textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \overline{x})^2] \\
&= \mathrm{E}(\textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2) - \mathrm{E}(\hat{\beta}_1^2) \sum_{i=1}^n (x_i - \overline{x})^2 \\
&= \mathrm{E}[\textstyle\sum_{i=1}^n (Y_i^2 - 2Y_i\overline{Y} + \overline{Y}^2)] - [\mathrm{Var}(\hat{\beta}_1) + (\mathrm{E}(\hat{\beta}_1))^2] \sum_{i=1}^n (x_i - \overline{x})^2
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}[\textstyle\sum_{i=1}^n Y_i^2 - 2n\overline{Y}^2 + n\overline{Y}^2] - [\frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2} + \beta_1^2] \textstyle\sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \textstyle\sum_{i=1}^n \mathrm{E}(Y_i^2) - n\mathrm{E}(\overline{Y}^2) - \sigma^2 - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \textstyle\sum_{i=1}^n [\mathrm{Var}(Y_i) + (\mathrm{E}(Y_i))^2] - n[\mathrm{Var}(\overline{Y}) + (\mathrm{E}(\overline{Y}))^2] - \sigma^2 - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= n\sigma^2 + \textstyle\sum_{i=1}^n (\beta_0+\beta_1 x_i)^2 - \sigma^2 - n(\beta_0+\beta_1\overline{x})^2 - \sigma^2 - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \sigma^2(n-2) + \textstyle\sum_{i=1}^n [(\beta_0+\beta_1 x_i)^2 - (\beta_0+\beta_1\overline{x})^2] - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \sigma^2(n-2) + \textstyle\sum_{i=1}^n (\beta_0^2 + \beta_1^2 x_i^2 + 2\beta_0\beta_1 x_i - \beta_0^2 - \beta_1^2\overline{x}^2 - 2\beta_0\beta_1\overline{x}) - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \sigma^2(n-2) + \beta_1^2 \textstyle\sum_{i=1}^n (x_i^2 - \overline{x}^2) + 2\beta_0\beta_1 \sum_{i=1}^n x_i - 2\beta_0\beta_1 \sum_{i=1}^n \overline{x} - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \sigma^2(n-2) + \beta_1^2 \textstyle\sum_{i=1}^n (x_i^2 - 2x_i\overline{x} + \overline{x}^2) - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \sigma^2(n-2) + \beta_1^2 \textstyle\sum_{i=1}^n (x_i-\overline{x})^2 - \beta_1^2 \sum_{i=1}^n (x_i-\overline{x})^2 \\
&= \sigma^2(n-2).
\end{aligned}
$$

Det følger dermed at $\mathrm{E}(s^2) = \frac{\mathrm{E(SSE)}}{n-2} = \sigma^2\frac{n-2}{n-2} = \sigma^2$. Dette er et generelt bevis når vi antar uavhengig, identisk fordelte støyledd, dvs vi antar ikke nødvendigvis normalfordeling. Resultatet er nok velkjent, men jeg har ikke funnet det i noen av kildene jeg har brukt.

4.2.2 Konfidensintervaller[14]

Hvis Z er en standardnormalfordelt stokastisk variabel, og V er kji-kvadratfordelt med v frihetsgrader, har vi at $T = \frac{Z}{\sqrt{\frac{V}{v}}}$ er en t-fordelt stokastisk variabel med v frihetsgrader. For mer om dette, se for eksempel [Walpole m.fl 2002], kapittel 8.7.

Det kan vises at for normalfordelte data er $s^2 = \frac{\mathrm{SSE}}{n-2}$ kji-kvadratfordelt med $n-2$ frihetsgrader, og uavhengig av $\hat{\beta}_1$. Vi har tidligere vist at $\hat{\beta}_0$ og $\hat{\beta}_1$ er forventningsrette estimatorer, og vi har vist hva variansene er. En forutsetning om normalfordelte støyledd gjør at $\hat{\beta}_0$ og $\hat{\beta}_1$ blir normalfordelte, siden begge disse estimatorene da vil være lineære funksjoner av uavhengige stokastiske normalfordelte variable.

Vi husker at $\mathrm{Var}(\hat{\beta}_1) = \sigma_{\beta_1}^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i-\overline{x})^2}$. Vi definerer $S_{xx} = \sum_{i=1}^n (x_i-\overline{x})^2$, og ved å bruke det vi vet om t-fordelingen får vi følgende t-fordelte observator: $T = \frac{(\hat{\beta}_1-\beta_1)/(\sigma/\sqrt{S_{xx}})}{s/\sigma} = \frac{\hat{\beta}_1-\beta_1}{s/\sqrt{S_{xx}}}$. Et $(1-\alpha)$ konfidensintervall for β_1 blir derfor:

$$\hat{\beta}_1 - \frac{t_{\alpha/2}\cdot s}{\sqrt{S_{xx}}} < \beta_1 < \hat{\beta}_1 + \frac{t_{\alpha/2}\cdot s}{\sqrt{S_{xx}}},$$

der vi finner $t_{\alpha/2}$ for eksempel i en t-fordelingstabell, ved $n-2$ frihetsgrader.

[14] Teorien i denne delseksjonen er hentet fra [Walpole m.fl 2002], kapittel 11.5-11.6.

Med $(1-\alpha)$ konfidensintervall mener vi altså at $(1-\alpha)$ er sannsynligheten for at den aktuelle parameteren befinner seg i det intervallet. Et tilsvarende resonnement som det for β_1 gir følgende $(1-\alpha)$ konfidensintervall for β_0:

$$\hat{\beta}_0 - \frac{t_{\alpha/2} \cdot s\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}} < \beta_0 < \hat{\beta}_0 + \frac{t_{\alpha/2} \cdot s\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}.$$

Konfidensintervall til én responsverdi (prediksjonsintervall)
Et tilsvarende resonnement gir oss følgende $(1-\alpha)$ prediksjonsintervall for én responsverdi, som vi her betegner med y_0:

$$\hat{y_0} - t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}} < y_0 < \hat{y_0} + t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}},$$

der x er en gitt verdi av forklaringsvariabelen, og $\hat{y_0}$ er den estimerte verdien av responsvariabelen, gitt x.

Kvantilregresjon overflødig i normaltilfellet
Når vi har normalfordeling, kan vi finne kvantilene på ethvert sted i responsfordelingen ved å lage prediksjonsintervall. Et $(1-\alpha)$ prediksjonsintervall vil da gi $\tau = \alpha/2$-kvantilen i venstre endepunkt, og $\tau = (1-\alpha/2)$-kvantilen i høyre endepunkt. Men når vi ikke har normalfordeling vil denne metoden kunne feile radikalt, og den bør da ikke brukes. Mer om dette på side 61, hvor vi sammenligner kvantiler regnet ut ved konfidensintervall med kvantilene regnet ut ved kvantilregresjon i forbindelse med et av eksemplene i seksjon 4.6, hvor vi ikke har normalfordeling.

4.2.3 Forutsetninger og kommentarer

Vi forutsetter altså i den ordinære regresjonsmodellen at vi har normalfordelte støyledd der variansen σ^2 er konstant, uavhengig av verdien til forklaringsvariabelen, noe som i mange situasjoner ikke holder. I situasjoner hvor forutsetningene ikke holder vil denne prosedyren kunne slå feil, noe som vil gi gale resultater i for eksempel hypotesetesting.

Denne gjennomsnittbaserte regresjonsmetoden har samme problem som gjennomsnitt, nemlig at det finnes ingen begrensning på hvor stor innflytelse uteliggere kan ha. Se seksjon 4.6 for eksempler på et større datasett. Ofte eliminerer man uteliggere fra datasettet, men i noen situasjoner er det meningen at det skal være en del uteliggere, så dette vil ikke gi et optimalt resultat.

Ellers har vi at om to fordelinger har samme forventningsverdi og samme standardavvik men ulik skjevhet, så vil formen kunne være svært forskjellig, noe gjennomsnittbasert regresjonsanalyse ikke greier å fange opp, men som man vil lykkes med i kvantilregresjon (se [Hao og Naiman 2007] side 24-25 for mer om dette).

4.3 Medianregresjon

Teorien i denne og neste seksjon er basert på [Hao og Naiman 2007], side 34-38. Eksemplene som er brukt for å illustrere teorien er egenproduserte.

Når vi har svært skjeve fordelinger, vil det kunne være vanskelig å gi en tolkning av forventningsverdien, eller tilsvarende tolkning av betinget gjennomsnitt i regresjonsanalyse. Metoden med betinget median i regresjonsanalyse vil kunne gi oss mer fornuftige sentralestimater. Vi vil i denne seksjonen ta for oss medianregresjon. Dette er et spesialtilfelle av kvantilregresjon, som er beskrevet i seksjon 4.4.

Vi ser på den enkle, lineære regresjonsmodellen som er beskrevet på side 29. I stedet for å minimere total kvadratavstand kan vi velge å minimere total absoluttavstand. Sett i lys av teorien i seksjon 3.2 vil vi nå få medianestimatorer av regresjonskoeffisientene.

For støyleddene i denne situasjonen antar vi ingen spesiell fordeling, bare at de har median lik 0, noe som gir oss den estimerte regresjonslinjen $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$. Den totale absoluttavstanden i y-retning, SAE (sum of absolute errors), skriver vi som:

$$\text{SAE} = \sum_{i=1}^{n} |y_i - \tilde{y}_i| = \sum_{i=1}^{n} |y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i|.$$

Når vi minimerer SAE med hensyn på $\tilde{\beta}_0$ og $\tilde{\beta}_1$ får vi en regresjonslinje som går gjennom to datapunkter og som har omtrent like mange datapunkt over linjen som under. Regresjonslinjen vil ligge i midten, tilsvarende slik medianen lå i midten da vi minimerte gjennomsnittlig absoluttavstand for en observasjonsmengde i seksjon 3.2.

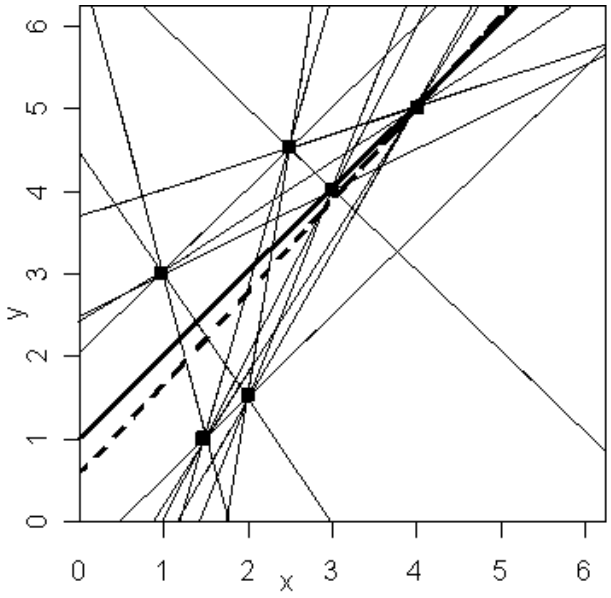
Mer presist har vi generelt for kvantilregresjon et krav[15] om at regresjonslinjen ligger slik at $\frac{N}{n} \leq \tau \leq \frac{N+2}{n}$, der n er antall observasjoner, N antall negative residualer (dvs antall punkter som ligger under den estimerte regresjonslinjen) og 2 markerer antall nullresidualer. En nullresidual er et punkt regresjonslinjen går gjennom, og i situasjonen vi har sett på er det 2 slike punkt. Tilsvarende har vi $\frac{P}{n} \leq 1 - \tau \leq \frac{P+2}{n}$, der P er antall positive residualer. Dette gjelder altså generelt for kvantilregresjon, i situasjonen for medianregresjon setter vi $\tau = 0.50$. Det vil si at når n er stor, vil vi ha enten like mange eller så å si like mange positive som negative residualer, og regresjonslinjen vil ligge i midten, eller tilnærmet i midten. Som regel er det flere linjer mellom par av datapunkter som vil oppfylle disse kravene, og løsningen som gir medianestimatet er den linjen som har de $\tilde{\beta}_0$ og $\tilde{\beta}_1$ som gir minste verdi av SAE. Det er ikke alltid vi har unik løsning. Mer om dette i Eksempel 4.3 på side 41.

Det er her altså snakk om hvor mange punkt som ligger over eller under regresjonslinjen, og ikke *hvor* langt over eller under punktene ligger. Dette viser en form for robusthet for kvantilregresjonslinjene, akkurat som tilfellet var for den empiriske medianen og de empiriske kvantilene i en tallmengde, som vi så i kapittel 2. Men det finnes situasjoner hvor denne robustheten i kvantilregresjon feiler. For ethvert

[15] For bevis og flere detaljer, se [Koenker 2005] kapittel 2.2.2 side 36-38.

datasett kan vi lage en uteliggende verdi som er slik at alle kvantilregresjonslinjer må ha dette punktet som et av de punktene linjene går gjennom. Vi kan fremdeles få en linje som oppfyller kravene på forrige side, med omtrent like mange positive som negative residualer. At man risikerer dette er ikke uventet, da man ville fått store utslag på total absoluttavstand om regresjonslinjen verken hadde passert gjennom eller i nærheten av en slik uteliggende observasjon. For at dette skal skje må altså ingen andre potensielle linjer mellom observasjonspar være i nærheten av denne uteliggeren. Vi vil se eksempler på noe tilsvarende i seksjon 4.6, i situasjoner hvor vi har regresjonskurver av andre og tredje grad, hvor tilsvarende problem gjelder. Dette er imidlertid ikke like alvorlig, da det bare er i utkanten, hvor kurvene uansett har høy varians, at det kan skje. For linjer vil det prege resultatet overalt, men selv om det er mulig i teorien, så er det ytterst sjeldent at dette skjer i praksis i det lineære tilfellet. I seksjon 4.6 vil vi se at dette er et problem som også oppstår for minimering av kvadratavstand. For flere detaljer omkring dette, se [Koenker 2005] side 44.

Eksempel 4.2
Vi ser på datasettet fra Eksempel 4.1,
$(x_i, y_i) = [(1,3); (1.5,1); (2,1.5); (2.5,4.5); (3,4); (4,5)]$, der Y er responsvariabelen og X er forklaringsvariabelen. På Figur 4.2 under har vi plottet de seks datapunktene, og laget en linje mellom hvert par. Linjen markert med fet strek er medianregresjonslinjen, noe vi skal vise i dette eksempelet. Stiplet linje er gjennomsnittsregresjonslinjen fra Eksempel 4.1, og vi ser at i dette tilfellet ligger de nokså nær hverandre.



Figur 4.2: *De seks datapunktene fra Eksempel 4.2, med linje mellom hvert par av punkter, der én av disse linjene vil være medianregresjonslinjen (markert med fet strek her, den stiplede linjen er gjennomsnittsregresjonslinjen fra Eksempel 4.1).*

Vi ser at vi kan ha både 1, 2 og 3 negative residualer (det samme for positive, men vi forholder oss videre til negative residualer), og vi vil likevel ha at kravet vi formulerte for antall negative residualer på side 36 er oppfylt. Det vil si at vi ikke nødvendigvis får en regresjonslinje med like mange datapunkter over og under når vi har så få observasjoner. Vi har 15 linjer, og vi kan ekskludere de fem potensielle regresjonslinjene som har null eller fire negative residualer, slik at vi står igjen med ti alternativer, og vi må sjekke hvilket alternativ det er som gir lavest verdi av SAE. I Tabell 4.1 under har vi satt opp hvilke av de ti aktuelle par av datapunkter som gir hvilke $\tilde{\beta}_0$ og $\tilde{\beta}_1$, og regnet ut hvilken verdi vi får for SAE i hver situasjon.

Datapunktpar i (x,y)-planet	$(\tilde{\beta}_0, \tilde{\beta}_1)$	SAE	Ant. neg. res.	Ant. pos. res.
$(1.5, 1)$ og $(3, 4)$	$(-2, 2)$	6	2	2
$(2, 1.5)$ og $(2.5, 4.5)$	$(-10.5, 6)$	23.5	2	2
$(1, 3)$ og $(3, 4)$	$(2.5, 0.5)$	5.5	2	2
$(3, 4)$ og $(4, 5)$	**(1,1)**	**5**	2	2
$(1.5, 1)$ og $(2.5, 4.5)$	$(-4.25, 3.5)$	12	3	1
$(1, 3)$ og $(4, 5)$	$(2\frac{2}{3}, \frac{2}{3})$	$5\frac{1}{3}$	3	1
$(2.5, 4.5)$ og $(3, 4)$	$(7, -1)$	13	3	1
$(1.5, 1)$ og $(4, 5)$	$(-1.4, 1.6)$	5.6	1	3
$(2, 1.5)$ og $(3, 4)$	$(-3.5, 2.5)$	8	1	3
$(1, 3)$ og $(2, 1.5)$	$(4.5, -1.5)$	15.5	1	3

Tabell 4.1: *Hvert datapunktpar som er listet opp tilsvarer en potensiell medianregresjonslinje, og den korrekte linjen er den som har lavest SAE (sum of absolute errors).*

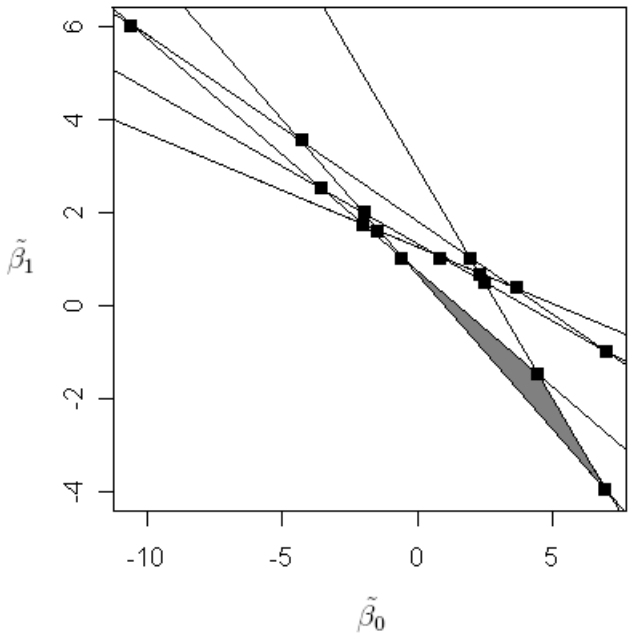
Vi ser at datapunktparet i den fjerde raden i tabellen gir minste verdi av SAE, og vi får altså at medianestimatene for regresjonskoeffisientene i dette tilfellet blir $\tilde{\beta}_0 = 1$ og $\tilde{\beta}_1 = 1$. Disse to estimatene gir den estimerte medianregresjonslinjen som er markert med fet linje i Figur 4.2 på forrige side, og vi ser at vi i denne situasjonen får like mange positive som negative residualer.

Situasjonen vi nå har sett på var enkel, men likevel tok det en del tid å regne seg fram til løsningen, langt mer tid enn vi brukte i gjennomsnittbasert regresjonsanalyse i Eksempel 1 på det samme datasettet. Men hva hvis vi har 1000 observasjoner i stedet for bare seks? Vi vil da få $\frac{1000 \cdot 999}{2} = 499\ 500$ linjer mellom par av observasjoner å jobbe med! Og det er når vi bare har to regresjonskoeffisienter. For p regresjonskoeffisienter og n observasjoner får vi $\binom{n}{p} = \frac{n!}{(n-p)!p!}$ linjer å jobbe med. Det vil være umulig i praksis å finne løsningene manuelt i løpet av rimelig tid når n og p er relativt store. Imidlertid er det blitt utviklet algoritmer (mer om dette på de neste sidene) som gjør at datamaskiner enkelt kan foreta disse utregningene for oss. Kort forklart

går én av flere mulige algoritmer ut på at man begynner med et par av observasjoner (som altså tilsvarer en linje hvor $(\tilde{\beta}_0, \tilde{\beta}_1)$ er henholdsvis konstantledd og stigningstall), og bytter ut dette paret med et nytt par bestående av en ny observasjon og en av de to foregående, og man velger den nye observasjonen som danner de $(\tilde{\beta}_0, \tilde{\beta}_1)$ som gjør at SAE blir minst mulig. Når det ikke går an å gjøre SAE mindre ved denne metoden, så har vi funnet de verdiene av $\tilde{\beta}_0$ og $\tilde{\beta}_1$ som gir medianestimatene til regresjonskoeffisientene. Se avsnittet under for flere detaljer.

Mer om algoritmen som leder fram til medianestimatorene

Vi ser på datasettet vi brukte i Eksempel 4.1 og Eksempel 4.2. Vi kan transformere grafen i Figur 4.2 til en ny graf hvor vi har $\tilde{\beta}_0$ som førsteakse og $\tilde{\beta}_1$ som andreakse. Se Figur 4.3 under. Hver av de 15 linjene i (x, y)-planet vil nå utgjøre et punkt i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet, og de seks datapunktene i (x, y)-planet vil danne seks linjer i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet hvor vi vil få ett punkt på hvert sted hvor linjene skjærer hverandre[16].

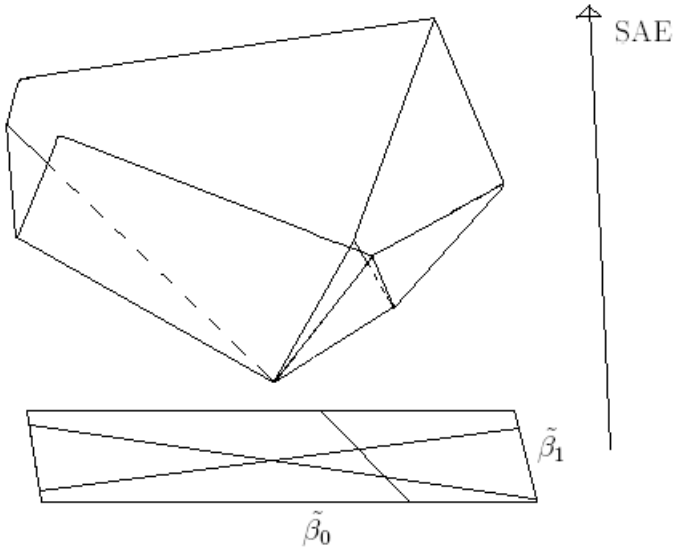


Figur 4.3: *Fra Eksempel 4.2, $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet, hvert av de 15 markerte punktene tilsvarer en mulig medianregresjonslinje.*

[16]Seks ikke-parallelle linjer gir $\frac{6 \cdot 5}{2} = 15$ skjæringspunkter.

Alle de fem punktene på en linje i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet har ett felles datapunkt i (x, y)-planet, og linjesegmentet mellom to punkt i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet tilsvarer to linjer fra (x, y)-planet som har ett felles datapunkt.

Tar vi for oss de mange mangekantregionene[17] dannet av de seks linjene i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet, får vi at hver region danner en familie linjer i (x, y)-planet. Alle linjene i hver slik familie deler punktene i datasettet (x_i, y_i) i to deler, slik at SAE, som vi skal minimere med hensyn på $\tilde{\beta}_0$ og $\tilde{\beta}_1$, vil være lineær i hver region. Lar vi så SAE utgjøre tredjeaksen i en graf med $\tilde{\beta}_0$ og $\tilde{\beta}_1$ som henholdsvis første- og andreakse, vil vi få en konveks polyedrisk overflate, hvor hvert hjørne tilsvarer et skjæringspunkt i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet, og dermed en linje i (x, y)-planet. Hver kant mellom to hjørner av overflaten tilsvarer et par slike linjer fra (x, y)-planet som har ett felles datapunkt. Se Figur 4.4 under for hvordan det vil se ut i et svært enkelt eksempel med bare tre datapunkter. Vi har tre innvendige hjørner i overflaten, der det ene hjørnet er et bunnpunkt[18].



Figur 4.4: *$(\tilde{\beta}_0, \tilde{\beta}_0)$-planet i bunn, hvert skjæringspunkt tilsvarer en mulig medianregresjonslinje, og tredjeaksen markerer SAE, slik at den polyedriske overflaten vil vise hvilket skjæringspunkt som gir bunnpunkt, og som dermed gir oss medianregresjonslinjen.*

[17] Et eksempel på en slik region er merket med grått i Figur 4.3 på forrige side.

[18] Det er mulig at en hel kant, eller en hel region får bunnverdien i stedet for bare et hjørne, men dette er svært sjelden og får dessuten ingen betydning i praksis, da alle løsninger vil være like gode.

Vi ser at denne situasjonen er tilnærmet lik den vi så da vi minimerte gjennomsnittlig absoluttavstand for en tallmengde i seksjon 3.2 (se Figur 3.3 på side 24), bare at vi nå minimerer for to parametre i stedet for én, slik at vi nå får en tredimensjonal figur. På Figur 3.3 har vi linearitet mellom hvert nabopar av datapunkter, og her (Figur 4.4 på forrige side, se også Figur 4.3) har vi linearitet mellom hvert nabopar av datapunktpar, det vil si hvert par av datapunktpar som har ett felles datapunkt. Hvert hjørne på Figur 4.3 og Figur 4.4 tilsvarer som nevnt ett datapunktpar, og mellom hvert datapunktpar som har ett felles datapunkt går det en kant. Felles for begge situasjonene er at vi får bunnpunkt i et hjørne, og at bunnpunktet ikke nødvendigvis er unikt, en kant eller en hel region kan også ligge i bunn, noe vi så tilsvarende for empirisk median i seksjon (3.2) og som vi vil se i Eksempel 4.3 under.

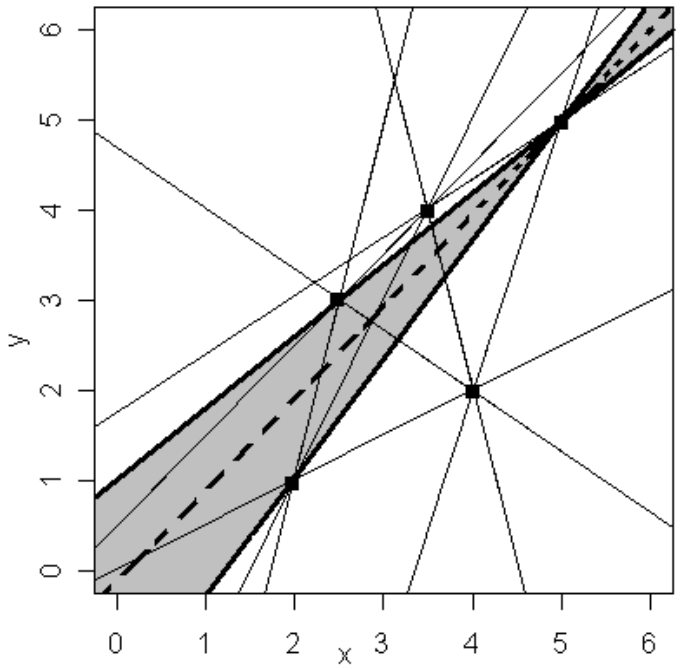
Vi kan nå lage en algoritme ved hjelp av lineærprogrammering (se side 46 for mer om dette). hvor vi begynner i et hvilket som helst hjørne, det vil si et hvilket som helst av skjæringspunktene i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet. En mulig algoritme[19] er å gå fra hjørne til hjørne av den polyedriske overflaten hvor vi hver gang velger å gå langs den kanten som fører oss til et nytt hjørne som gir lavest verdi av SAE, helt til vi ikke lenger kan finne et nytt hjørne som gir lavere mulig verdi av SAE, det vil si når alle retningsderiverte i SAE-retning i punktet er ikke-negative. Da vil vi være i bunnhjørnet (er dog som nevnt ikke sikkert løsningen er unik), nemlig det punktet i $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet som gir minste verdi av SAE, og fra dette bunnpunktet får vi de verdiene sv $\tilde{\beta}_0$ og $\tilde{\beta}_1$ som gir medianestimater til regresjonskoeffisientene. Alt dette kan så generaliseres til kvantilregresjon, hvor vi erstatter SAE med SWAE, som er total vektet absoluttavstand, definert på side 44.

Eksempel 4.3
På samme måte som at vi har to observasjoner i midten i et sortert datasett med jevnt antall observasjoner, der begge disse observasjonene minimerer total eller gjennomsnittlig absoluttavstand (det spiller ingen rolle hvilken av de to man ser på i denne sammenhengen), kan vi ha to regresjonslinjer som ligger i "midten" som begge minimerer SAE. At linjene ligger i midten betyr at total absoluttavstand i y-retning mellom linjene og datapunktene er den samme for de to linjene. Vi kan også ha tre regresjonslinjer som ligger i midten i en enkel, lineær regresjonsmodell. Vi vil da ha at en hel region danner bunnpunkt i den polyedriske overflaten vi så et eksempel på i Figur 4.4 på forrige side. Dette i stedet for bare en kant, som er tilfellet når vi har to linjer i midten. Dette er imidlertid noe som bare skjer når vi jobber med diskrete verdier, og jo flere observasjoner man har, jo sjeldnere vil det naturlig nok være. I praksis, når n er stor nok, vil det skje med sannsynlighet tilnærmet lik 0. Skulle det likevel skje vil det ikke få noen spesielle konsekvenser, da alle løsninger

[19] En alternativ, mer effektiv algoritme er å identifisere hvilken kant som går brattest nedover fra punktet vi er i, altså går vi videre langs den kanten som har lavest retningsderivert i SAE-retning, og videre til neste punkt. Dette krever mindre regnekraft enn å regne ut SAE for alle nabopunkter.

er likeverdige, og er n stor vil løsningene uansett ligge svært nær hverandre. For en detaljert drøfting omkring dette, se [Koenker 2005] side 34-38.

Et enkelt eksempel som gir oss to løsninger er datasettet:
$(x_i, y_i) = [(2,1); (2.5,3); (3.5,4); (4,2); (5,5)]$, der Y er responsvariabelen og X er forklaringsvariabelen. Av ti mulige datapunktpar er det fem som oppfyller kravene for residualer som nevnt på side 36, nemlig de fem som gir en potensiell regresjonslinje med enten én negativ residual og to positive, eller omvendt.



Figur 4.5: *De fem datapunktene fra Eksempel 4.3, med linje mellom hvert par av punkter, der linjen som gir de $\tilde{\beta}_0$ og $\tilde{\beta}_1$ som gir lavest verdi av SAE vil være medianregresjonslinjen (her to likeverdige linjer, markert med fet linje, og midtlinjen mellom disse to, den stiplede linjen, kan vi kalle den sanne medianregresjonslinjen).*

Datapunktpar i (x,y)-planet	$(\tilde{\beta}_0, \tilde{\beta}_1)$	SAE
$(2.5,3)$ og $(4,2)$	$(4.67, -0.67)$	$7\frac{2}{3}$
$(2,1)$ og $(3.5,4)$	$(-3,2)$	6
$(2.5,3)$ og $(5,5)$	$(1,0.8)$	4
$(2,1)$ og $(5,5)$	$(-1\frac{2}{3}, 1\frac{1}{3})$	4
$(3.5,4)$ og $(4,2)$	$(18,-4)$	21

Tabell 4.2: *Hvert datapunktpar som er listet opp tilsvarer en mulig medianregresjonslinje, og den korrekte linjen er den som har lavest SAE (sum of absolute errors). Her er det to likeverdige linjer som har bunnverdien av SAE.*

Vi ser i Tabell 4.2 på forrige side at vi får to regresjonslinjer som minimerer SAE, disse er markert med fet linje i Figur 4.5 på samme side. Vi ser at de gir nokså ulikt resultat, men dette er et resultat av at det her kun er snakk om fem datapunkter. Midt mellom de to linjene er det markert en fet, stiplet linje som vi kan kalle den "sanne" medianregresjonslinjen, mer om dette i neste avsnitt. Denne linjen vil altså også minimere total absoluttavstand, det vil også hele familien linjer som er merket av i grått i Figur 4.5 på forrige side gjøre, slik at det ikke er slik at en linje *må* gå gjennom to datapunkter for å minimere absoluttavstand, men algoritmene vil kun finne løsninger som går gjennom to datapunkter.

I statistikkpakken `R` (mer om hvordan man tilpasser kvantilregresjonsmodeller i `R` i seksjon 4.5 på side 49) får vi bare oppgitt én løsning selv om det er flere, vi får da bare en advarsel om at det kan være at løsningen ikke er unik. Dette tyder på at algoritmen ser at det finnes minst én retningsderivert i bunnpunktet som er lik 0, noe som betyr at det finnes en kant til et nytt hjørne som også må ha bunnverdien. Sett i lys av teorien for empirisk median til en endelig, sortert mengde observasjoner, hvor denne er definert til å være midtpunktet mellom de to observasjonene i midten når vi har jevnt antall observasjoner, så vil den korrekte løsningen av $\tilde{\beta}_0$ og $\tilde{\beta}_1$ være midtpunktet på kanten som forbinder de to alternativene når vi ser på $(\tilde{\beta}_0, \tilde{\beta}_1)$-planet. I (x, y)-planet får vi da en linje som går gjennom det datapunktet de to alternative løsningene hadde til felles, og midt mellom de to andre involverte punktene, som sammen med det ene fellespunktet utgjorde de to løsningene. I Figur 4.5 på forrige side er denne midtpunktlinjen i (x, y)-planet merket med fet, stiplet linje, og dette vil jeg da definere til å være den sanne medianregresjonslinjen i dette tilfellet. I teorien som er utviklet snakker man ikke om slike midtpunktlinjer, bare om at vi kan få flere løsninger.

At algoritmen ikke dekker denne situasjonen kan være fordi dette er såpass sjelden, spesielt i virkelige situasjoner med stor n, at det ikke vil oppstå, og skulle det oppstå så vil de to alternative linjene ligge såpass nær hverandre at det utgjør liten forskjell, akkurat slik de to observasjonene i midten av en sortert mengde med jevnt antall observasjoner sannsynligvis vil ligge svært nær hverandre når n er stor. I tillegg, har vi p parametere kan vi ha $p + 1$ løsninger (da det er dimensjonen vi da jobber i, for en hel region kan ligge i bunn), og finner man ut at man har en løsning som ikke er unik må man også lete seg fram til alle de andre løsningene og så regne ut midtpunktet, og det man tjener på dette i praksis, hvor dette så å si aldri skjer og hvor de ulike linjene uansett vil ligge svært nær hverandre, er så lite at man burde kunne sløyfe det.

Vi har nå sett på medianregresjon for en enkel, lineær regresjonsmodell med to parametere, hvor det å finne estimatorene til regresjonskoeffisientene gikk ut på å finne bunnpunktet til en polyedrisk, tredimensjonal overflate. Dette generaliseres for multiple modeller til at vi får $p+1$ dimensjoner, når vi har p parametere vi minimerer med hensyn på. De samme prinsippene vil gjelde for all form for medianregresjon.

4.4 Kvantilregresjon

Hittil i dette kapitlet har vi sett på to ulike måter å utføre regresjonsanalyse på hvor vi finner estimatorer som estimerer sentralitet. Hva om vi ikke er så interessert i å bare estimere størrelser som karakteriserer sentrum, men heller kan tenke oss å også analysere hva som vil skje nærmere den ene eller den andre halen? For eksempel se på hvor mye skatt de 5 % rikeste betaler, eller hvordan inntekter utvikler seg for de 10 % av befolkningen med lavest inntekt, betinget på utvalgte forklaringsvariable? En måte å kunne se på ting som dette, er å generalisere medianregresjon til å gjelde for alle kvantiler, slik at vi selv kan velge hvilken kvantil vi vil basere analysen på. Det er dette som er kvantilregresjon, første gang omtalt av Koenker og Basset i 1978.

Siden vi kan velge å se på hvilken kvantil vi vil, kan vi også se på mange kvantiler samtidig. For eksempel kan vi estimere én regresjonslinje for hver femte prosentil, det vil si betingede estimater for $Q^{(0.05)}, Q^{(0.10)}, ..., Q^{(0.95)}$, altså 19 forskjellige linjer, hvor vi i tillegg til å se på hvordan hver linje isolert sett utvikler seg kan analysere hvordan linjene forandrer seg i forhold til hverandre, betinget på ulike verdier av forklaringsvariable. Da vil 0.50-kvantilen vise forandringer sentralt, mens de andre kvantilene vil vise forandringer i form og spredning, og påvise eventuell skjevhet.

Vi ser fremdeles på den enkle, lineære regresjonsmodellen $Y|x = \beta_0 + \beta_1 x + \varepsilon$, men her skal vi ha $Q^{(\tau)}(Y|x) = \beta_0^{(\tau)} + \beta_1^{(\tau)} x$, slik at for hver verdi av τ har vi et støyledd som vi ikke antar noen spesiell fordeling for, men som er lik 0 for τ-kvantilen[20]. Vi definerer altså forskjellige støyledd for hver τ-verdi, slik at det ikke er snakk om ett støyledd hvis τ-kvantil er lik 0 for alle τ. Estimater for $\beta_0^{(\tau)}$ og $\beta_1^{(\tau)}$ kan finnes ved å minimere total vektet absoluttavstand, SWAE (sum of weighted absolute errors), med hensyn på de to regresjonskoeffisientene. Dette blir tilsvarende slik vi beskrev at vi kunne gjøre for å finne estimerte kvantiler i en tallmengde i seksjon 3.3:

$$\text{SWAE} = (1-\tau) \sum_{y_i < \hat{\beta}_0^{(\tau)} + \hat{\beta}_1^{(\tau)} x_i} |y_i - \hat{\beta}_0^{(\tau)} - \hat{\beta}_1^{(\tau)} x_i| + \tau \sum_{y_i > \hat{\beta}_0^{(\tau)} + \hat{\beta}_1^{(\tau)} x_i} |y_i - \hat{\beta}_0^{(\tau)} - \hat{\beta}_1^{(\tau)} x_i|$$

Det paret av $\hat{\beta}_0^{(\tau)}$ og $\hat{\beta}_1^{(\tau)}$ som minimerer SWAE gir oss τ-kvantilestimater av regresjonskoeffisientene. Dette gjøres tilsvarende slik vi beskrev for spesialtilfellet medianen i forrige seksjon. Vi skal ikke gi en lignende detaljert beskrivelse her, men beskrive den manuelle framgangsmåten i et enkelt eksempel. I tillegg skal vi kort omtale lineærprogrammering, som danner grunnlaget for algoritmene som brukes i statistisk programvare for å tilpasse modeller til datasett i kvantilregresjon (se side 46). De

[20] En forutsetning som kan oppfattes som noe ulogisk for τ-verdier nær 0 eller 1, noe som kan sies å forsterke det vi senere vil si om at resultane i kvantilregresjon blir mindre sikre jo lenger vi fjerner oss fra $\tau = 0.50$ i den ene eller den andre retningen.

estimerte kvantilregresjonslinjene vi får vil ha omtrent 100τ % av observasjonene under linjen og $100(1-\tau)$ % over linjen[21]. Dette er et resultat av vektingen i minimaliseringsproblemet, der datapunktene som ligger under linjen blir vektet med τ og de over blir vektet med $(1-\tau)$. Har vi få datapunkter får vi relativt få mulige regresjonslinjer, slik at en linje kan dekke et intervall av kvantiler, se Eksempel 4.4 under. Kvantilregresjonslinjer vil alltid dekke et intervall kvantiler, men for stor n vil intervallene som regel bli svært korte.

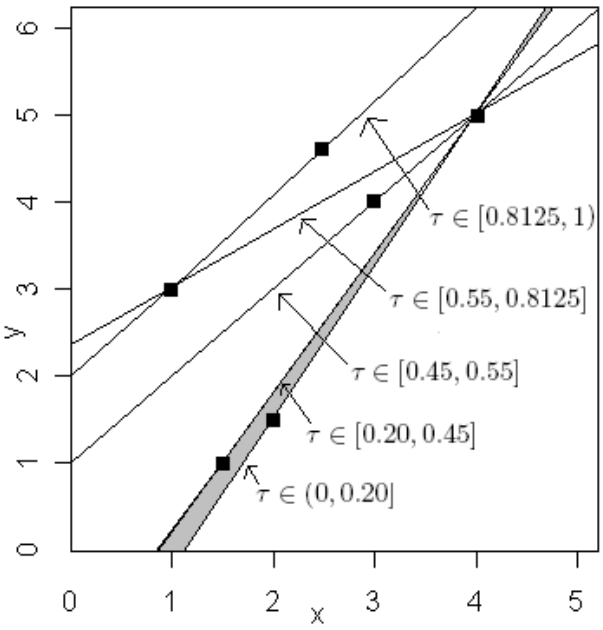
Eksempel 4.4
Ser vi på datasettet fra Eksempel 4.2 på side 37, hvor vi fant medianestimater for β_0 og β_1, så kan vi tilsvarende finne kvantilestimater. Vi hadde altså datasettet $(x_i, y_i) = [(1,3);(1.5,1);(2,1.5);(2.5,4.5);(3,4);(4,5)]$. Vi skal nå finne $\hat{\beta}_0^{(0.20)}$ og $\hat{\beta}_1^{(0.20)}$. I Tabell 4.3 under ser vi på hvilket datapunktpar som minimerer total vektet absoluttavstand SWAE, der vi setter $\tau = 0.20$ i formelen for SWAE på forrige side. I følge regelen for antall residualer over og under regresjonslinjen, kan vi ha maksimalt 1 negativ residual. Dette betyr at det er kun de seks linjene som gir enten 0 eller 1 negativ residual som er interessante i denne sammenhengen, og vi vil derfor kun se på disse i tabellen under.

Datapunktpar i (x,y)-planet	$(\hat{\beta}_0^{(0.20)}, \hat{\beta}_1^{(0.20)})$	SWAE	Neg. res.	Pos. res.
$(1.5,1)$ og $(2,1.5)$	$(-0.5,1)$	1.6	0	4
$(2,1.5)$ og $(4,5)$	**(-2,1.75)**	**1.3**	0	4
$(1,3)$ og $(1.5,1)$	$(7,-4)$	6.6	0	4
$(1.5,1)$ og $(4,5)$	**(-1.4,1.6)**	**1.3**	1	3
$(2,1.5)$ og $(3,4)$	$(-3.5,2.5)$	2.5	1	3
$(1,3)$ og $(2,1.5)$	$(4.5,-1.5)$	3.85	1	3

Tabell 4.3: *Hvert datapunktpar som er listet opp tilsvarer en potensiell kvantilregresjonslinje, og den korrekte linjen er den som har lavest SWAE (sum of weighted absolute errors).*

Vi ser at vi her får to løsninger, og tilsvarende som for medianregresjon i Eksempel 4.3 på side 41 kunne vi definert en linje midt i mellom de to likeverdige løsningene til å være den sanne linjen. Som i Eksempel 4.3 vil hele regionen som er merket av med grått i Figur 4.6 på neste side være løsninger av minimaliseringsproblemet. Men `R` gir oss bare én løsning, med advarsel om at det kan være at løsningen ikke er unik. I dette tilfellet kommer algoritmen i `R` fram til $(\hat{\beta}_0^{(0.20)}, \hat{\beta}_1^{(0.20)}) = (-2, 1.75)$. Vi bruker nå `R` til å finne kvantilregresjonslinjer fra $\tau = 0.05$, $\tau = 0.10$ og så videre til $\tau = 0.95$. Resultatet ser vi i Figur 4.6 under.

[21] Mer presist om antall datapunkter over og under en kvantilregresjonslinje, se side 36.



Figur 4.6: *Kvantilregresjonslinjer for datasettet i Eksempel 4.4, og hvilket τ-intervall hver linje dekker. Grå region minimerer SWAE for $\tau = 0.20$.*

Vi ser at vi her har kun fem linjer som dekker hele spekteret av kvantiler. Videre ser vi at vi har fire kvantiler hvor vi får to løsninger (noe som er naturlig i og med at vi har fem linjer, da vi får fire overganger mellom linjer). Det dreier seg om kvantilene $\tau = 0.20$ (som vi alt har oppdaget på forrige side), $\tau = 0.45$, $\tau = 0.55$ og $\tau = 0.8125$. Har man langt flere observasjoner vil vi som regel få langt flere linjer, hvor hver linje dekker et mye mindre τ-intervall. I de situasjoner vi da får mer enn én løsning vil det få liten betydning, i og med at da vil de to linjene ligge svært tett inntil hverandre, slik at resultatet blir temmelig likt uansett hvilket av de to alternativene som blir brukt.

Vi ser ellers på figuren at linjen algoritmen i `R` kom fram til har 0.20-kvantilen som øvre endepunkt av det kvantilintervallet den dekker. Imidlertid er det ingen fast regel for algoritmen at den slik går nedenfra og oppover hva kvantiler angår, dette ser vi om vi for eksempel vil finne 0.45-kvantilen, hvor vi får løsningen $(\hat{\beta}_0^{(0.20)}, \hat{\beta}_1^{(0.20)}) = (1, 1)$ i `R`, altså linjen som har 0.45-kvantilen som nedre endepunkt. Slik vi har beskrevet mulige algoritmer ser vi at hvilken løsning en algoritme ender opp med avhenger av hvilket punkt den velger å starte i, se bl.a. om lineærprogrammering under.

Lineærprogrammering

Problemer som søker å optimere en lineær funksjon som har lineære restriksjoner kalles for *lineære programmer*. Problemet med å minimere SWAE kan skrives som et slikt lineærprogram (dette er hentet fra [Koenker 2005] side 7):

$$\min_{(\beta_0,\beta_1,u,v)\in\mathbb{R}^2\times\mathbb{R}_+^{2n}} \{\tau 1_n^\top u + (1-\tau)1_n^\top v | \beta_0 + \beta_1 x + u - v = y\}.$$

Her står n for antall observasjoner, x og y er de n-dimensjonale vektorene med observasjonene av henholdsvis forklaringsvariabelen og responsvariabelen og $1_n^\top$ er en n-dimensjonal vektor med 1-tall. Betingelsen kan også skrives som $y-\beta_0-\beta_1 x = u-v$, og positiv og negativ del ved denne residualvektoren er definert ved henholdsvis u og v, som begge er n-dimensjonale vektorer, der

$$u_i = \left\{ \begin{array}{cc} y_i-\beta_0-\beta_1 x_i, & y_i-\beta_0-\beta_1 x_i > 0 \\ 0, & y_i-\beta_0-\beta_1 x_i \leq 0 \end{array} \right\} \text{ og } v_i = \left\{ \begin{array}{cc} 0, & y_i-\beta_0-\beta_1 x_i \geq 0 \\ -(y_i-\beta_0-\beta_1 x_i), & y_i-\beta_0-\beta_1 x_i < 0 \end{array} \right\}.$$

Den positive delen i det som skal minimeres blir vektet med τ og den negative med $1-\tau$, på samme måte som vi så i seksjon 3.3 på side 27. Vi har en lineær funksjon betinget på en polyedrisk mengde (den lineære betingelsen sett i sammenheng med mengden $\mathbb{R}^2 \times \mathbb{R}_+^{2n}$ som gir $(2n+2)$-dimensjonalitet) som skal minimeres. En lineærprogrammeringsalgoritme[22] vil finne bunnpunktet tilsvarende slik det er skissert for tilfellet med medianen i seksjon 4.3. Man begynner i et hjørne av den polyedriske overflaten (det spiller ingen rolle hvilket, men forskjellige starthjørner kan gi forskjellige løsninger i de situasjoner løsningen ikke er unik), og algoritmen finner den kanten som går brattest nedover, det vil si som har lavest retningsderivert, og man går fra hjørne til hjørne på den måten til man når bunnen, for da finnes det ikke lenger noen negativ retningsderivert. Alternativt kan man lage en algoritme hvor man fra et starthjørne går videre til det nabohjørnet som gir lavest SWAE, men det viser seg at dette ikke er like effektivt.

4.4.1 Standardavvik til estimatorene til regresjonskoeffisientene

Teorien i denne seksjonen er hentet fra [Hao og Naiman 2007], side 44-50.

Vi definerer her X til å være observasjonsmatrisen, med dimensjon $n\times(p+1)$. Rad nummer i inneholder 1 i første kolonne, og videre verdiene av alle de p forklaringsvariablene for observasjon nummer i, dvs vi antar at β_0 er et konstantledd. Antar vi identisk, uavhengig fordelte støyledd, får vi følgende kovariansmatrise[23] for $\vec{\hat{\beta}}(\tau)$ (vektor som inneholder alle de $p+1$ regresjonskoeffisientene):

$$\Sigma_{\vec{\hat{\beta}}(\tau)} = \frac{\tau(1-\tau)}{n} \cdot \frac{1}{f_\varepsilon(0)^2}(X^\top X)^{-1}.$$

[22]For detaljer om matematikken bak, se [Koenker 2005] kapittel 6 (og liten introduksjon om emnet på sidene 5-10 i samme bok).

[23]Bygger på det faktum at fordelingen til $\hat{Q}^{(\tau)}$ for stor n er tilnærmet normalfordelt med forventning $Q^{(\tau)}$ og varians $\frac{\tau(1-\tau)}{n} \cdot \frac{1}{f(Q^{(\tau)})^2}$, se [Hao og Naiman 2007] side 11.

På diagonalen til denne kovariansmatrisen har vi variansen til hvert element av $\vec{\hat{\beta}}(\tau)$, den første øverst til venstre og videre nedover mot høyre. Et estimat for standardavviket til den k'te regresjonskoeffisienten, $s_{\hat{\beta}_k^{(\tau)}}$, finner vi ved å ta kvadratroten av den tilhørende variansen. Konfidensintervall får vi ved formelen $\hat{\beta}_k^{(\tau)} \pm z_{\alpha/2} s_{\hat{\beta}_k^{(\tau)}}$, som altså er en normaltilnærming. Leddet f_ε er tettheten til støyleddene. Vi er interessert i tettheten i punktet 0 fordi i følge definisjonen for støyleddene på side 44 har vi at vi antar τ-kvantilen til støyleddene er nettopp 0. Men siden vi ikke kjenner $f_{\varepsilon^{(\tau)}}$, så må den estimeres. Vi kan estimere $\frac{1}{f_{\varepsilon^{(\tau)}}} = \frac{d}{d\tau} Q^{(\tau)}$ ved $\frac{1}{2h}(\hat{Q}^{(\tau)}(\tau + h) - \hat{Q}^{(\tau)}(\tau - h))$. Valg av båndbredden h er et kapittel for seg selv, se [Koenker 2005], side 139-140 eller [Wasserman 2006], kapittel 6 for en beskrivelse av det.

Dersom vi ikke antar uavhengig, identisk fordelte støyledd, så har ikke lenger støyleddene samme tetthet, så vi må vekte hver observasjon med tilhørende estimert tetthet av støyleddet i punktet 0 (antakelsen $\varepsilon^{(\tau)} = 0$ gjelder fortsatt), slik at vi får:

$$\Sigma_{\vec{\hat{\beta}}^{(\tau)}} = \frac{\tau(1-\tau)}{n} D_1^{-1} D_0 D_1^{-1}, \text{ der } D_0 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x^{(i)\top} x^{(i)}, \text{ og } D_1 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} w_i x^{(i)\top} x^{(i)},$$

der $x^{(i)}$ er rekke i av matrisen X og $w_i = f_{\varepsilon_i}(0)$, og disse vektene må estimeres.

Bootstrapmetoden

Et alternativ når man skal finne standardavvik til regresjonskoeffisientene er bootstrapmetoden, hvor man ikke antar noen spesiell fordeling for responsvariabelen. Denne metoden ble introdusert av Efron i 1979.

Vi har opprinnelig et estimat $\hat{\beta}_k^{(\tau)}$. Vi bruker det opprinnelige datasettet med n observasjoner, og trekker et nytt datasett *med tilbakelegging* fra dette, også det på n observasjoner. Enkelte observasjoner kan altså trekkes flere ganger, mens andre kan bli ekskludert. Man repeterer denne prosedyren M ganger, hvor M vanligvis er mellom 50 og 200 for estimering av standardavvik, og mellom 500 og 2000 for estimering av konfidensintervall.

For hvert av de M bootstrap-datasettene estimerer vi kvantilestimatet vi er ute etter, slik at vi får $\hat{\beta}_{k,m}^{(\tau)}$, der $m = 1, 2, ..., M$. Vi får altså til sammen M kvantilestimater, og vi finner standardavviket til disse på vanlig måte:

$$s_{boot} = \sqrt{\frac{1}{M-1} \sum_{m=1}^{M} (\hat{\beta}_{k,m}^{(\tau)} - \frac{1}{M} \sum_{i=1}^{M} \hat{\beta}_{k,i}^{(\tau)})^2}.$$

Dette bootstrap-standardavviket er et estimat for standardavviket til $\hat{\beta}_k^{(\tau)}$, altså til det opprinnelige estimatet av $\beta_k^{(\tau)}$. Vi kan nå for eksempel lage et $(1-\alpha)$ konfidensintervall for $\beta_k^{(\tau)}$ ved formelen $\hat{\beta}_k^{(\tau)} \pm z_{\alpha/2} s_{boot}$, som altså er en normaltilnærming.

Det finnes mange forskjellige metoder innen bootstrapping utover det vi har skissert her. For mer om dette (utover [Hao og Naiman 2007] side 47-50, som er brukt som kilde her), se for eksempel [Efron 1979] og [Wasserman 2006], kapittel 3.

4.5 Kvantilregresjon i R

Funksjonsrutiner basert på lineærprogrammering som gir kvantilregresjonsestimater er tilgjengelig i for eksempel det statistiske programmet R, som er et åpent og gratis statistisk programsystem. Vi skal her først og fremst se på det som trengs for å utføre kvantilregresjon i eksemplene i seksjon 4.6. For detaljer henvises det til [Koenker 2005], Appendix A.

Den statistiske programvaren R kan lastes ned via http://www.r-project.org/. Det er blitt laget en statistisk pakke til R, som blant annet inneholder verktøy til å finne kvantilregresjonslinjer. Denne kan lastes ned via Internett ved å skrive følgende på kommandolinjen i R:

```
> install.packages("quantreg")
```

Installering skal skje automatisk så lenge man er logget på Internett. For å gjøre pakken tilgjengelig for den nåværende kjøringen av R, skriv:

```
> library(quantreg)
```

Denne siste kommandoen må man skrive hver gang man kjører R for å kunne bruke pakken.

Vi skal konsentrere oss om funksjonsrutinen `rq`, og hvordan den brukes til å finne kvantilregresjonslinjer. Vi antar at funksjonsrutinen for å tilpasse minste kvadraters modell, `lm`, er kjent, og vi husker i så fall at den brukes slik for en enkel, lineær regresjonsmodell:

`> lm(y~x)`, der `y` er vektoren med observerte responsverdier og `x` er vektoren med observerte verdier av forklaringsvariabelen. Rutinen `rq` fungerer på samme måte, men vi må her spesifisere hvilken kvantil vi vil se på, for eksempel:

`>rq(y~x, tau=0.50)`, i dette tilfellet velger vi $\tau = 0.50$, som betyr at vi får median-estimater for β_0 og β_1. Vi velger altså `tau` til å være den kvantilen vi er interessert i.

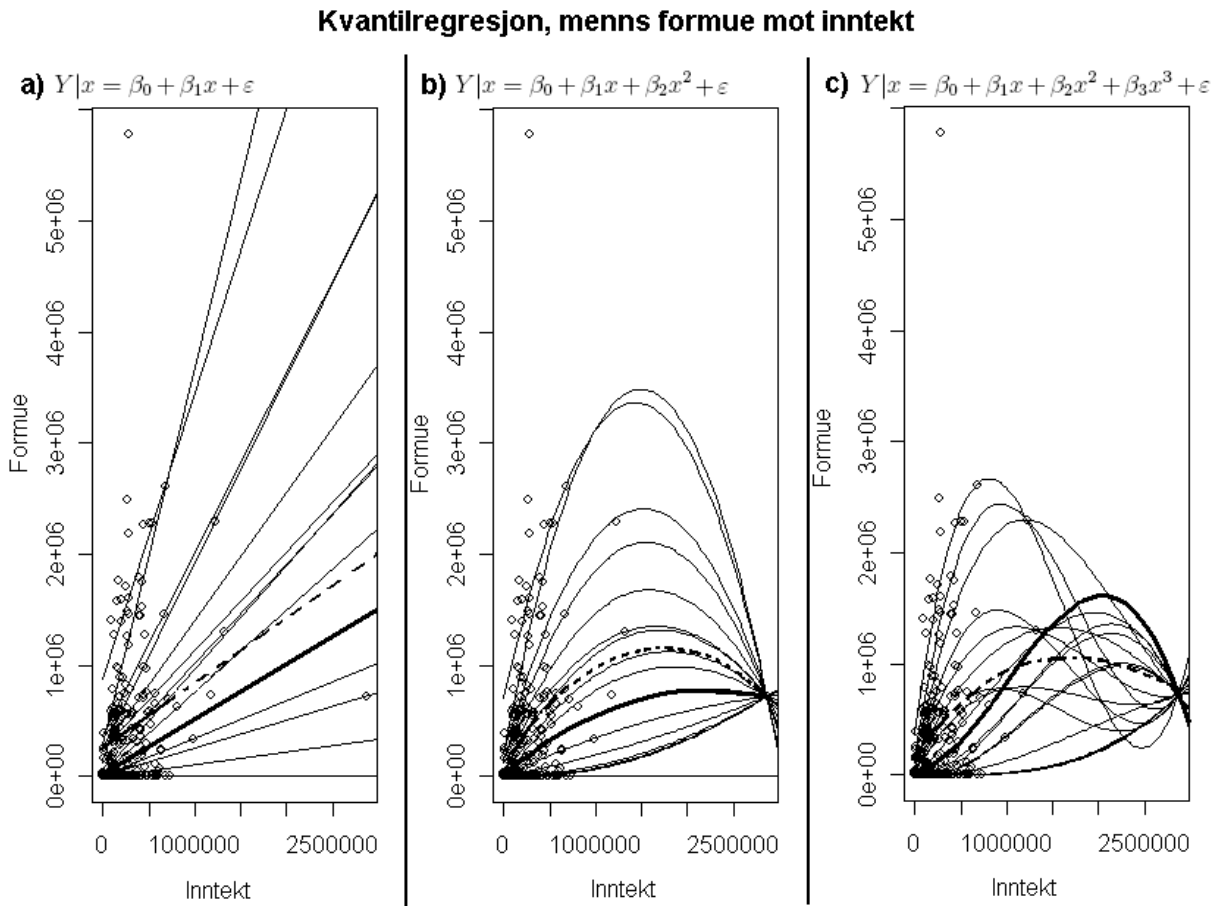
I denne pakken finnes det også en funksjonsrutine for ikke-trivielle ikke-lineære modeller, `nlrq`, som fungerer på samme måte som funksjonsrutinen for ikke-trivielle ikke-lineære minste kvadraters modeller, `nls`. Trivielle ikke-lineære modeller, som for eksempel inkludering av andregradsledd som vi skal se på i neste seksjon, tilpasses ved hjelp av `rq`, på samme måte som ved `lm`.

4.6 Eksempler: Kvantilregresjonsanalyse på datasett

Vi returnerer til skattelistedatasettet vi tok for oss i eksempelet i seksjon 2.7, som står opplistet i appendikset, seksjon A.1. Vi tar for oss menn og kvinner hver for seg, og foretar enkel, lineær regresjonsanalyse, som beskrevet i seksjon 4.1, med formue som responsvariabel og inntekt som forklaringsvariabel. Vi skal se på både kvantilregresjon og ordinær minste kvadraters regresjon, og sammenligne disse. Situasjonen er nå så omfattende at vi bruker `R` til å hjelpe oss, slik vi forklarte i seksjon 4.5. I tillegg vil vi demonstrere ikke-lineær kvantilregresjon ved å ha med to plott hvor vi utvider den enkle, lineære modellen slik at vi får modeller på formen $Y|x = \beta_0+\beta_1 x+\beta_2 x^2+\varepsilon$ og $Y|x = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$. Prinsippene bak multippel kvantilregresjon, når vi har $p+1$ regresjonskoeffisienter, er akkurat som for den enkle varianten. Der skulle vi finne hvilket par av regresjonskoeffisienter som minimerte total vektet absoluttavstand (se side 44), og dette generaliseres til at vi må finne hvilken trio, eller hvilken kvartett, og så videre, som minimerer total vektet absoluttavstand. At vi har andre- og tredjegradsledd i situasjonen vi ser på her, gjør det ikke noe ekstra vanskelig, da det bare er å skrive $z = x^2$ og $w = x^3$, og dermed kan modellene behandles som om de er lineære.

Noe som er verdt å diskutere før vi setter gang, er den spesielle situasjonen som kommer som følge av hvordan formue (Y er responsvariabelen, i dette tilfellet formue) er definert. Vanligvis når vi har kontinuerlig fordeling så har vi $P(Y = y) = 0$, men slik formue er definert vil alle verdier som i realiteten er negative få verdien 0, slik at vi får $P(Y = 0) > 0$. For mennenes analyse, for eksempel, får vi $P(\widehat{Y = 0}) = \frac{90}{241}$, altså et estimat for sannsynligheten ved å ta antall observasjoner med formue lik 0 delt på totalt antall observasjoner. Hvilke konsekvenser får dette når vi skal gjøre regresjonsanalyse? For de lave kvantilene vil det føre til at mange linjer/kurver vil bli en vannrett strek gjennom $y = 0$. Dette gjelder stort sett opp til og med $\tau = 0.30$. Dette betyr ikke annet enn at ved 0.30-kvantilen og lavere, så viser analysen at formue vil være 0 uansett inntekt, og i og med at vi har så mange observasjoner med formue lik 0, så er ikke dette uventet. Det vil ikke få noen konsekvenser for de øvrige kvantilregresjonskurvene, for som vi har sett før, så er det kun antall observasjoner som ligger under og over linjene som spiller inn, og om de som ligger under ligger på $y = 0$ eller har negative verdier spiller ikke inn, så lenge vi befinner oss i kjerneområdet. For gjennomsnittet derimot er det klart at vi får mye større verdier enn vi ville fått om de reelle, negative verdiene skulle stått i stedet for 0. Og det er klart at grunnen til at vi her får en kraftig høyreskjevhet er nettopp fordi vi har definert 0 til å være laveste mulige verdi. Men i og med at formue er definert på denne måten må vi også forholde oss til det.

Vi begynner med mennenes analyse, se Figur 4.7 på neste side, og videre Figur 4.8, hvor den samme analysen er foretatt, bare at to svært uteliggende observasjoner er fjernet. Plottene blir så drøftet hver for seg og videre sammenlignet.



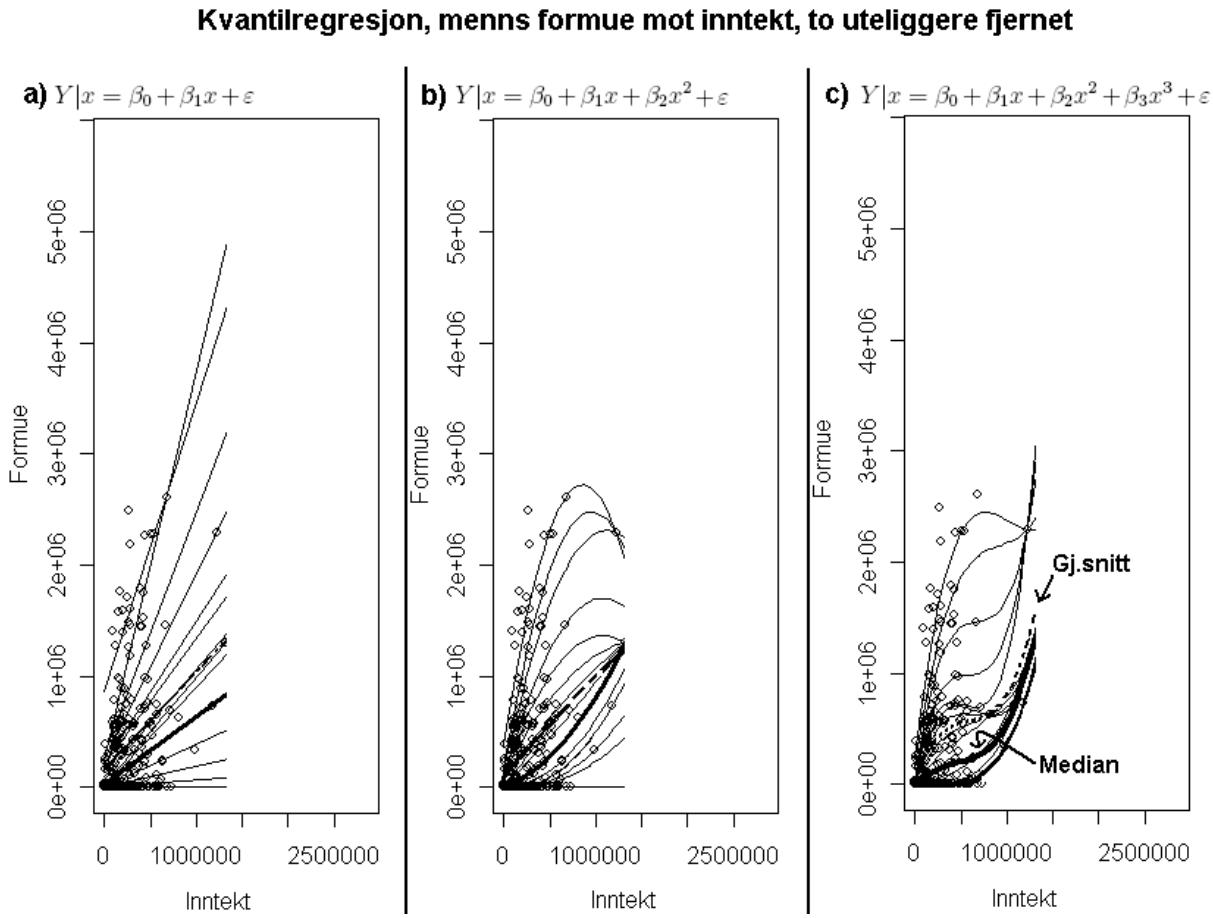
Figur 4.7*: Menns formue er respons og tilhørende inntekt er forklaringsvariabel, 241 observasjoner. 19 kvantilregresjonskurver*[24] *er tegnet inn, for kvantilene* $\tau = 0.05, \tau = 0.10, ..., \tau = 0.95$*. De seks første i modell a går oppå hverandre av grunner nevnt på side 50, dvs vi har* $(\hat{\beta}_0^{(\tau)}, \hat{\beta}_1^{(\tau)}) = (0, 0)$ *for* $\tau \in (0, 0.30)$*. Vi har også noen oppå hverandre nederst i b og c. I tillegg er minste kvadraters regresjonskurven tegnet inn, med stiplet kurve. Medianregresjonskurven (*$\tau = 0.50$*) er markert med fet kurve. Kurvene er i stigende rekkefølge fra* $\tau = 0.05$ *og oppover i kjerneområdet*[25] *til observasjonene. I utkanten*[26] *forekommer det krysninger av kurvene.*

Vi legger merke til at i modell b og c, så blir kurvene presset gjennom et punkt langt ute til høyre, og dette punktet er altså en uteliggende observasjon, se side 36 for mer om dette fenomenet.

[24] *Kurve* er her en fellesbetegnelse på linjer og kurver som ikke er rette.

[25] Når vi videre snakker om *kjerneområdet,* menes det de verdiene av forklaringsvariabelen hvor de fleste observasjonene er.

[26] Når vi videre snakker om *utkanten,* menes det området utenfor kjerneområdet til forklaringsvariabelen.



Figur 4.8: *Menns formue er respons og tilhørende inntekt er forklaringsvariabel, 239 observasjoner (to uteliggere fjernet). 19 kvantilregresjonskurver er tegnet inn, for kvantilene* $\tau = 0.05, \tau = 0.10, ..., \tau = 0.95$. *De seks første i modell a går oppå hverandre av grunner nevnt på side 50, dvs vi har* $(\hat{\beta}_0^{(\tau)}, \hat{\beta}_1^{(\tau)}) = (0, 0)$ *for* $\tau \in (0, 0.30)$. *Også noen oppå hverandre nederst i b og c. I tillegg har vi tegnet inn minste kvadraters regresjonskurven, med stiplet kurve. Medianregresjonskurven (*$\tau = 0.50$*) er markert med fet kurve. Kurvene er i stigende rekkefølge fra* $\tau = 0.05$ *og oppover i kjerneområdet til observasjonene. I utkanten forekommer det krysninger av kurvene. Plottene er i samme skala som i Figur 4.7 for at vi skal kunne sammenligne direkte, men vi har ikke tatt med kurvene til høyre for det som blir den ytterste observasjonen når uteliggerne er fjernet, da resultatet utenfor observasjonenes rekkevidde ikke er interessant i denne sammenhengen.*

Vi ser først på analysen vist i Figur 4.7. Vi begynner med den lineære modellen, til venstre i figuren. Her ser vi at enkelte av kvantilregresjonslinjene krysser hverandre. Vi ser at dette skjer langt ute, i områder hvor det ikke finnes noen observasjoner. Dette er en logisk feil, men noe som er uunngåelig bare vi beveger oss langt nok ut til den ene eller den andre siden. Vi siterer [Koenker og Xiao 2006]:

"[...]linear quantile regression applications [...] should be cautiously interpreted as useful local approximations to more complex nonlinear global models. If we take the linear form of the model too literally then obviously at some point, or points, there will be "crossings" of the conditional quantile functions – unless these functions are precisely parallel in which case we are back to the pure location shift form of the model."

Ellers observerer vi som ventet at gjennomsnittsregresjonslinjen ligger over median-regresjonslinjen, noe som tyder på høyreskjevhet[27]. Forskjellen er betydelig, linjene er nesten parallelle og siden differansen mellom skjæringspunktene med y-aksen er så forskjellige, er den relative forskjellen for lave inntektsverdier stor. Ellers er det større spredning i den øvre halvdelen av kvantilregresjonslinjene enn i den nedre, noe som igjen tyder på høyreskjevhet.

Går vi over på de ikke-lineære analysene, i midten og til høyre i figuren, ser vi at alle kurvene samsvarer svært bra med den lineære modellen i kjerneområdet, der hvor det er flest observasjoner. Men går vi videre ut til høyre, ser vi at alle kurvene, også minste kvadraters kurven, blir presset gjennom ett punkt, og dette punktet er en svært uteliggende observasjon (mer om dette fenomenet på side 36). Denne observasjonen, som har svært høy inntekt og til sammenligning svært lav formue, setter sterke føringer på det som skjer når inntekt blir høy, i og med at det ikke er noen andre observasjoner med i nærheten av så høy inntektsverdi. Så i følge informasjonen vi sitter på i dette datasettet, så er resultatet logisk, men vi vet at det nok ikke er vanlig at høy inntekt skal gi lav formue, slik at i forhold til virkeligheten gir ikke dette noe bra bilde. Generelt er det slik at regresjonsmodeller er best i det området man har flest observasjoner, og det er nok begrenset nytteverdi, det vil si knapt noen overhodet, i informasjonen vi får i et område langt ute hvor det bare er én observasjon. Jo flere observasjoner, jo mindre varians vil resultatet naturlig nok ha. Fra [Hao og Naiman 2007] på side 11-12, har vi som tidligere nevnt at den empiriske kvantilen $\hat{Q}^{(\tau)}$ har forventning $Q^{(\tau)}$ og varians $\frac{\tau(1-\tau)}{n} \cdot \frac{1}{f(Q^{(\tau)})^2}$. Med andre ord er variansen omvendt proporsjonal til antall observasjoner i et område, og et område med kun én observasjon gir dermed relativt høy varians. I så måte vil vi her stole mest på den lineære modellen når vi er langt ute, for kvantilregresjonslinjene blir ikke påvirket av uteliggerne der (med unntak av i helt spesielle situasjoner, se

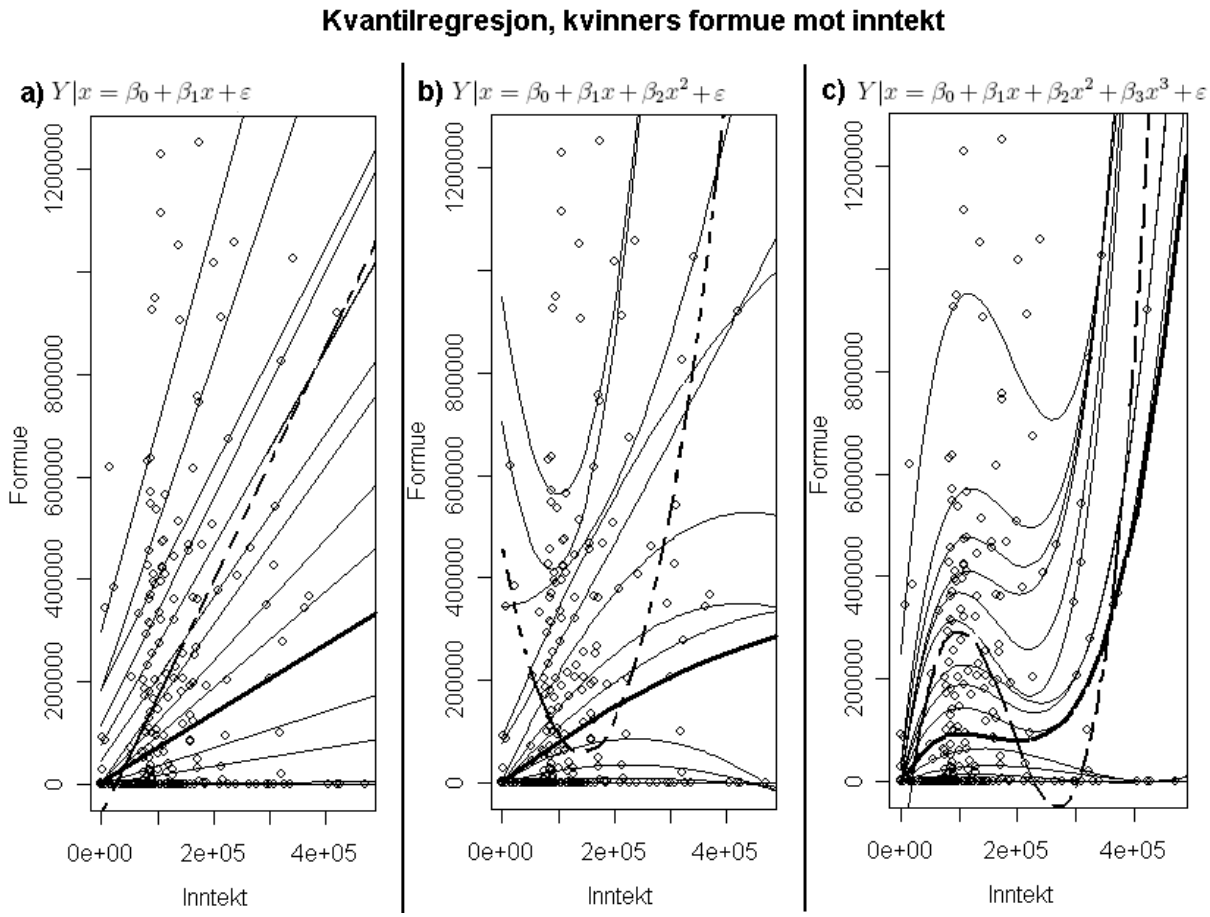
[27] Se side 16 for definisjon av høyreskjevhet.

side 36), men vi husker imidlertid at også den modellen har svakheter der ute siden linjene krysser hverandre. Generelt er det vanskelig å bruke regresjonsanalyse, både minste kvadraters og den kvantilbaserte metoden til å si noe sikkert om situasjonen i et område i utkanten hvor det er ingen eller svært få observasjoner.

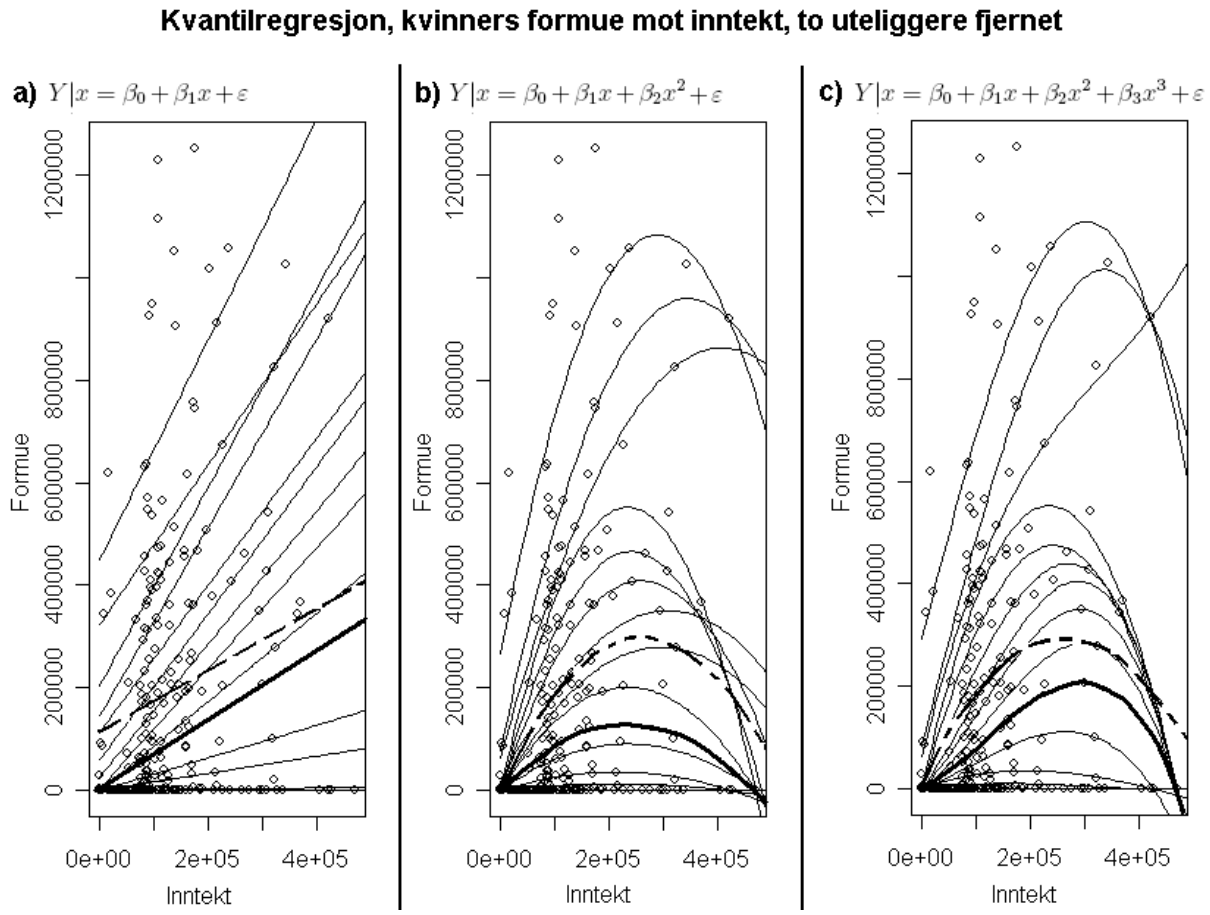
Vi ser så på Figur 4.8, hvor vi har foretatt den samme analysen som i Figur 4.7, bare at to svært uteliggende observasjonene er fjernet. Her ser vi at det ikke er så store forskjeller i kjerneområdet, noe som kan ha en sammenheng med at de to uteliggerne som er fjernet er i hver sin retning (en med høy inntekt i forhold til formue, og en med høy formue i forhold til inntektsverdi) og dermed nøytraliserer hverandre noe. Vi kan ikke ut fra denne analysen konkludere med at minste kvadraters regresjonskurver er mindre robust enn medianregresjonskurver, noe som vi vil se at vi mer kan hevde for kvinnenes analyse senere, hvor vi har to svært uteliggende observasjoner i samme retning.

Videre ser vi på kvinnenes analyse, vist i Figur 4.9 og Figur 4.10 på de følgende sider. Vi gjør som for mennenes analyse, bare at her har vi fjernet to svært uteliggende observasjoner fra plottene, for at vi bedre skal få fokusert på kjerneområdet. Disse to uteliggende observasjonene er tatt med i utregningene i Figur 4.9, men ikke i Figur 4.10 (dvs alle observasjoner som er med i utrekningen er med på Figur 4.10). De to det dreier seg om er henholdsvis [formue 9 037 636 kr, inntekt 576 154 kr] og [formue 2 300 409 kr, inntekt 25 030 kr], og vi ser på Figur 4.9 på neste side at vi der ikke går høyere enn formue lik 1 250 000 kr, som er en øvre grense når de to nevnte uteliggerne er tatt bort.

Vi ser først på analysen vist i Figur 4.9 på neste side. I den lineære modellen ser vi at gjennomsnittslinjen begynner godt under 0 og krysser mange av kvantilregresjonslinjene. Mer om dette i diskusjonen som følger figurene på de neste sidene. Vi får ellers rimelige resultat for de aller fleste kurvene, men gjennomsnittskurven samt kurvene for henholdsvis $\tau = 0.90$ og $\tau = 0.95$ i andregradsmodellen i midten får en rar form, som kommer av den svært uteliggende observasjonen med svært høy formue og til sammenligning svært lav inntekt. Sammenligner vi mediankurven med gjennomsnittskurven her, så virker førstnevnte kurve å være et fornuftig sentralestimat, mens sistnevnte har tatt en svært merkelig form, noe som skyldes de to uteliggerne. Mediankurven er ikke mye forskjellig i de tre modellene, noe som tyder på at denne er svært robust, ikke bare for uteliggere men også for modellskifte. De resterende kvantilkurvene virker også fornuftige. Unntaket er de to øverste, og det er klart at også kvantilregresjonskurver, for τ nær 0 eller 1, lar seg påvirke av uteliggere. Dette siden kurvene befinner seg i et område som grunnet få observasjoner vil vi få svært stor varians, siden resultatet kunne blitt fullstendig annerledes om vi for eksempel hadde byttet ut den verste uteliggeren med en annen uteligger et helt annet sted.



Figur 4.9: *Kvinners formue er respons og tilhørende inntekt er forklaringsvariabel, 234 observasjoner(de to observasjonene med høyest formue (uteliggere) er ikke med på figuren, men er tatt med ved utregningene). 19 kvantilregresjonskurver er tegnet inn, for kvantilene* $\tau = 0.05, \tau = 0.10, ..., \tau = 0.95$. *De seks første i modell a går oppå hverandre av grunner nevnt på side 50, dvs vi har* $(\hat{\beta}_0^{(\tau)}, \hat{\beta}_1^{(\tau)}) = (0, 0)$ *for* $\tau \in (0, 0.30)$. *Også noen oppå hverandre nederst i b og c. I tillegg har vi tegnet inn minste kvadraters regresjonskurven, med stiplet kurve. Medianregresjonskurven (*$\tau = 0.50$*) er markert med fet kurve. Kurvene er i stigende rekkefølge fra* $\tau = 0.05$ *og oppover i kjerneområdet til observasjonene.*



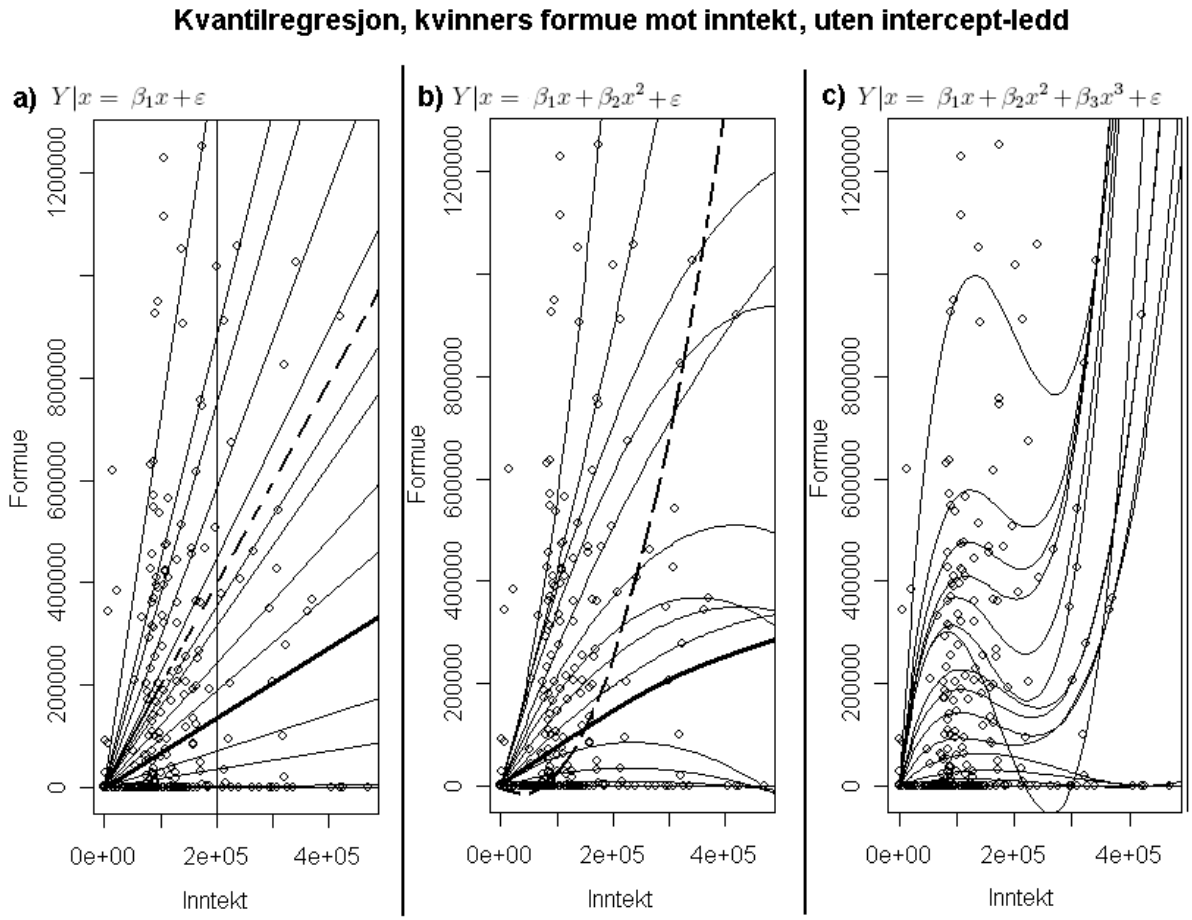
Figur 4.10: *Kvinners formue er respons og tilhørende inntekt er forklaringsvariabel, 232 observasjoner(de to observasjonene med høyest formue (uteliggere) er ikke med på figuren, og heller ikke tatt med ved utregningene). 19 kvantilregresjonskurver er tegnet inn, for kvantilene* $\tau = 0.05, \tau = 0.10, ..., \tau = 0.95$. *De seks første i modell a går oppå hverandre av grunner nevnt på side 50, dvs vi har* $(\hat{\beta_0}^{(\tau)}, \hat{\beta_1}^{(\tau)}) = (0, 0)$ *for* $\tau \in (0, 0.30)$. *Også noen oppå hverandre nederst i b og c. I tillegg har vi tegnet inn minste kvadraters regresjonskurven, med stiplet kurve. Medianregresjonskurven (*$\tau = 0.50$*) er markert med fet kurve. Kurvene er i stigende rekkefølge fra* $\tau = 0.05$ *og oppover i kjerneområdet til observasjonene.*

Ser vi på Figur 9 og Figur 10 og sammenligner, så ser vi at mediankurvene i kjerneområdet holder seg på omtrent samme sted. Gjennomsnittskurvene derimot forandrer seg en god del. Spesielt stor er forskjellen mellom situasjonen med og uten de to uteliggerne i den lineære modellen og andregradsmodellen. I dette tilfellet har begge uteliggerne relativt høye formueverdier i forhold til inntektsverdier, og vi ser da hvordan minste kvadraters regresjon er lite robust i forhold til dette, mens for kvantilregresjon er det små forskjeller med og uten uteliggere i kjerneområdet. Et lite unntak er det, når vi ser på τ oppimot 1, som vi ser når vi sammenligner andregradsmodellene her, av grunner tidligere nevnt.

Vi ser videre på analysen vist i Figur 4.9. I den lineære modellen til venstre ser vi at gjennomsnittsregresjonslinjen krysser svært mange av kvantilregresjonslinjene. Dette tyder på at denne linjen har latt seg påvirke i stor grad av uteliggere. Vi ser at stigningstallet til gjennomsnittsregresjonslinjen omtrent er tre ganger så stort som tilsvarende for medianregresjonslinjen, noe som tyder på en klar høyreskjevhet, noe vi også ser ved at spredningen mellom kvantilregresjonslinjen er større i øvre hale. Uteliggerne, som er tatt med ved utregningene i denne undersøkelsen, gjør at den estimerte responsvariabelen (estimert formue) ved gjennomsnittsregresjon gitt inntekt lik 0 blir -59 840. Ting bærer altså galt av sted for lave inntektsverdier, da formue her er definert til å være positiv. I tillegg ser vi for andregradsmodellen i midten at vi får en merkelig gjennomsnittskurve som har en relativt høy formueverdi for inntekt lik 0, men som synker kraftig før den stiger igjen. Dette problemet kan løses ved å innføre en forutsetning om at når inntekt er 0 er også formue lik 0 (det vil si sløyfe konstantleddet β_0), som vil gitt et noe annet resultat. Vi har foretatt en slik analyse[28] på kvinnenes datasett, se Figur 4.11 på neste side. Her har vi ellers samme situasjon som i Figur 4.9, det vil si de to svært uteliggende observasjonene er med i utregningene men ikke på plottet. Fordelen med denne omformuleringen er altså først og fremst at vi slipper negative formuer på lavt inntektsnivå, men betingelsen at formue skal være 0 når inntekt er 0 er streng, siden det ikke er noe i veien for at man kan ha å ha høy formue når inntekt er 0. Spesielt i høyre hale, for τ oppimot 1, så er det svært ulogisk å forutsette at formue skal være 0 når inntekt er 0, slik at det er en klar svakhet med å gjøre det på denne måten.

I Figur 4.10, hvor de to svært uteliggende observasjonene fra kvinnenes datasett er ekskludert fra utregningene, så ser vi at gjennomsnittskurvene i alle tilfeller ligger godt over mediankurvene, slik at det virker klart at vi har en betydelig høyreskjevhet i dataene.

[28] Dette kunne blitt problematisk i og med at R bare finner kvantilregresjonslinjer som går gjennom to datapunkter, for nå må de gjennom origo også, og å finne tre slike punkter på linje er i noen situasjoner umulig, men her går det siden vi har flere observasjoner som ligger nettopp i origo.



Figur 4.11: *Den samme analysen som i Figur 4.9, altså kvantilregresjon på datasettet med kvinnene fra side 19 hvor formue er respons og tilhørende inntekt er forklaringsvariabel. Forskjellen er at vi her krever at alle regresjonskurvene starter i punktet (0,0). Igjen er medianregresjonskurvene markert med fet linje, og gjennomsnittsregresjonskurvene markert med stiplet linje, og de andre kurvene utgjør som i de forrige figurene de resterende kvantilregresjonskurvene med et intervall på 0.05.*

Vi har 234 observasjoner, de to observasjonene med høyest formue (uteliggere) er ikke med på figuren, men er tatt med ved utregningene.

Vi ser først på den lineære analysen, til venstre i Figur 4.11 på forrige side. Vi kan her sammenligne betinget gjennomsnitt og betinget median direkte, siden begge begynner i punktet (0,0). Stigningstallet til minste kvadraters regresjonslinjen er 1.98, mens stigningstallet til medianregresjonslinjen er 0.69. Dermed gir altså betinget gjennomsnitt her en verdi som er nesten 3 ganger stor som betinget median! Gjennomsnittsregresjonslinjen ligger over kvantilregresjonslinjen for $\tau = 0.70$. Dette viser hvordan resultatene kan bli svært ulike for disse to sentralmålene når vi har skjeve fordelinger, og vi argumenterer da for at medianregresjonslinjen gir et bedre bilde av et typisk resultat midt i fordelingen, da størrelsen på uteliggere ikke virker inn.

For plottet i midten med andregradsleddet ser vi igjen at resultatene er nokså like som i den lineære modellen, men langt ute til høyre, hvor det er få observasjoner, får vi resultater som har begrenset nytteverdi, da variansen som nevnt vil være svært stor. Gjennomsnittsregresjonskurven blir her påvirket av uteliggere, noe som gjør at resultatet der det er flest observasjoner blir merkelig. Det er en forskjell det er verdt å merke seg. Altså at kvantilregresjonskurvene her ikke lar seg påvirke av uteliggere annet enn i områdene hvor uteliggerne er. Gjennomsnittsregresjonskurven kan derimot få ulogiske verdier overalt som følge av uteliggere, noe vi også ser for tredjegradsmodellen.

Når vi sammenligner denne undersøkelsen med tilsvarende undersøkelse hvor vi beholder konstantleddet (se Figur 4.9), så ser vi at det er nesten ingen endringer, med unntak av for de laveste inntektsverdiene. Vi bryr oss derfor ikke om å ta en tilsvarende undersøkelse hvor man fjerner uteliggerne.

Hva betyr betinget fordeling?

Vi sier at vi har betinget gjennomsnitt, betinget median og betingede kvantiler, og at de er betinget på forklaringsvariabelen. Det vil si at når vi har gitt en verdi av forklaringsvariabelen, så får vi disse betingede verdiene av responsvariabelen. I Figur 4.11a på forrige side har vi trukket en loddrett linje ved inntekt lik 200 000 kr, og vi ser at det er mye større spredning over medianen enn under, altså har vi høyreskjevhet. I Tabell 4.4 på neste side er verdiene til den betingede responsfordelingen listet opp for $\tau = 0.05$, $\tau = 0.10$,...,$\tau = 0.95$, gitt inntekt lik 200 000 kr.

$\hat{Q}^{(0.05)}(Y\|x = 200\ 000) =$	0
$\hat{Q}^{(0.10)}(Y\|x = 200\ 000) =$	0
$\hat{Q}^{(0.15)}(Y\|x = 200\ 000) =$	0
$\hat{Q}^{(0.20)}(Y\|x = 200\ 000) =$	0
$\hat{Q}^{(0.25)}(Y\|x = 200\ 000) =$	0
$\hat{Q}^{(0.30)}(Y\|x = 200\ 000) =$	0
$\hat{Q}^{(0.35)}(Y\|x = 200\ 000) =$	2 740.467
$\hat{Q}^{(0.40)}(Y\|x = 200\ 000) =$	36 030.3
$\hat{Q}^{(0.45)}(Y\|x = 200\ 000) =$	71 234
$\hat{Q}^{(0.50)}(Y\|x = 200\ 000) =$	138 016
$\hat{Q}^{(0.55)}(Y\|x = 200\ 000) =$	187 808.7
$\hat{Q}^{(0.60)}(Y\|x = 200\ 000) =$	241 579.6
$\hat{Q}^{(0.65)}(Y\|x = 200\ 000) =$	313 679.6
$\hat{Q}^{(0.70)}(Y\|x = 200\ 000) =$	352 213.8
$\hat{Q}^{(0.75)}(Y\|x = 200\ 000) =$	445 087.4
$\hat{Q}^{(0.80)}(Y\|x = 200\ 000) =$	581 065.8
$\hat{Q}^{(0.85)}(Y\|x = 200\ 000) =$	744 873.6
$\hat{Q}^{(0.90)}(Y\|x = 200\ 000) =$	872 739.2
$\hat{Q}^{(0.95)}(Y\|x = 200\ 000) =$	1 434 014
$\hat{\mathrm{E}}(Y\|x = 200\ 000) =$	396 200

Tabell 4.4: *Kvantilverdier og forventningsverdi for formue i kroner, gitt inntekt lik 200 000 kroner i eksempelet for kvinner uten konstantledd som vist i Figur 4.11.*

For tetthetsestimering kan man bruke en metode beskrevet i [Hao og Naiman 2007] side 11-12, hvor man kommer fram til formelen:

$$1/f(Q^{(\tau)}) = \frac{d}{dp}Q(\tau) \approx \frac{1}{2h}(\hat{Q}^{(\tau+h)} - \hat{Q}^{(\tau-h)}),$$

og dermed kan altså en tilnærming for tettheten på ulike steder regnes ut. Man har tatt utgangspunkt i en graf med kvantilene på førsteaksen og kvantilfunksjonene på andreaksen. Dette er ikke annet enn en graf som viser fordelingsfunksjonen, og siden vi vet at $f = F'$, så får vi tilnærmelsen over. Valg av båndbredden h for de ulike kvantilene i en slik situasjon er et mye omtalt tema i statistikk, som vi ikke skal gå nærmere inn på her. Se for eksempel [Koenker 2005] side 139-140, eller [Wasserman 2006], kapittel 6 for mer om tetthetsestimering.

Konfidensintervall kan bli ubrukelig uten normalitet
Vi husker teorien om konfidensintervall fra seksjon 4.2.2. Vi vil nå vise ut fra dette eksempelet hvordan denne teorien kan gi gale resultat når normalitetsforutsetningen ikke er oppfylt. Vi tar for oss inntekt på 200 000 kr for kvinnene, og vi vil under se på et 0.90 konfidensintervall, hvis endepunkter skal gi 0.05-kvantilen og 0.95-kvantilen dersom vi har normalfordeling. Alt er regnet ut i `R`.

Vi kaller nå responsvariabelen for y og forklaringsvariabelen for x. For kvinner i analysen fra Figur 4.11 har vi $s = \sqrt{\text{SSE}/233} = 616\ 800$, $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = 2.138906 \cdot 10^{12}$, $\overline{x} = 130\ 953.4$ og vi skulle ha $x = 200\ 000$. Den estimerte verdien av formue gitt inntekt lik 200 000 blir $\hat{y}|_{x=200\ 000} = 396\ 200$. Ved 233 frihetsgrader[29] vil vi få $t_{0.05} \approx z_{0.05} = 1.645$. Bruker vi formelen for prediksjonsintervall for $y_0|x$ fra seksjon 4.2.2, får vi følgende 0.90 konfidensintervall for $y_0|_{x=200\ 000}$:

$$396200 \pm 1.645 \cdot 616800\sqrt{1 + \tfrac{1}{234} + \tfrac{(200\ 000 - 130\ 953.4)^2}{2.138906 \cdot 10^{12}}} = 396\ 200 \pm 517\ 876.6,$$

som avrundet til nærmeste heltall gir endepunktene $(-121\ 676.6, 914\ 076.6)$. Ut fra henholdsvis 0.05- og 0.95-kvantilregresjonslinjen utregnet i `R` får vi for inntekt på 200 000 verdiene 0 og 1 434 014, som vi vil hevde er et bedre resultat. Resultatene i disse to tilfellene blir altså svært ulike, og det er naturlig å hevde at man kan få problematiske resultat når man bruker konfidensintervall til å finne ut hva kvantilene er når vi har skjeve fordelinger, som her.

4.7 Konklusjon

Konklusjonen blir at fordelen med kvantilregresjon er at vi får et bilde av hele fordelingen til responsvariabelen, gitt en forklaringsvariabel. Kvantilregresjon deler ellers svakheten til minste kvadraters regresjon ved at resultat i utkanten hvor det er få observasjoner har for stor varians til at de kan brukes til å si noe sikkert. Problemet med høy varians oppstår også for kvantilregresjonskurvene for kvantiler nær 0 eller 1, da det naturlig nok er relativt få observasjoner der. Analysen viser i tillegg at medianregresjonskurvene er svært så robuste for uteliggere i kjerneområdet, mens gjennomsnittsregresjonskurven i noen tilfeller kan forandre seg radikalt overalt, også i kjerneområdet, om man bare fjerner to observasjoner fra et datasett på knapt 250 observasjoner. Kvantilregresjonskurvene er robuste i kjerneområdet[30] til en forklaringsvariabel, men ikke for uteliggende verdier av en forklaringsvariabel.

Vi kan ellers konkludere med at innføring av modeller med andre- og tredjegradsledd ikke var spesielt fruktbart for eksemplene i seksjon 4.6, da vi stort sett fikk noenlunde samme resultater som i det lineære tilfellet i kjerneområdet, samt inkonsistente resultater utenfor.

[29] Vi har $n - 1 = 233$ frihetsgrader i denne situasjonen, siden vi estimerer kun én parameter.

[30] Med unntak av som flere ganger nevnt i helt spesielle situasjoner, som beskrevet på side 36.

5 Anvendelse på tidsrekkeanalyse

Vi skal i dette kapitlet forklare hvordan prinsippene bak kvantilregresjon kan overføres til andre områder hvor regresjon spiller inn. For eksempel finnes det muligheter for dette innen overlevelses- og tidsrekkeanalyse. Vi skal her se på tidsrekker. For bruk av kvantilregresjon innen overlevelsesanalyse, se for eksempel [Peng og Huang 2008], eller [Koenker 2005] side 250-255. Felles for disse temaene er at man nettopp er kommet i gang med å benytte seg av kvantilregresjonsmetoder, og eventuelle ulemper og problemer er ikke blitt skikkelig kartlagt ennå. Forskerne er enige om at mer arbeid trengs på dette feltet. For tidsrekker kommer dette fram i tidsskriftartikkelen fra 2006 av Koenker og Xiao, hvor kvantil autoregresjon blir introdusert. Artikkelen har et diskusjonstillegg, med fem diskusjonsinnlegg hvor andre forskere kommer med sine kommentarer til det som blir introdusert, før de opprinnelige artikkelforfatterne til slutt kommer med et tilsvar. Fei og Fei konkluderer i sin kommentarartikkel etter å ha diskutert mulige svakheter med at "*further studies are needed*", hvorpå Koenker og Xiao til slutt i sin kommentar til kommentarene vedgår at:

"*QAR models allow us to explore some features of time series that are inaccessible through classical methods, but they suffer from their own limitations, many of which have been brought out in this fruitful discussion. Much room is left for improvement; we look forward to the continuing discussion*".

Vi skal se på hvordan vi kan overføre kvantilregresjon til tidsrekkeanalyse i det enkleste tilfellet, og vi vil komme inn på hva det er som kan være en svakhet, som man ikke kommer bort i på samme måte ved vanlig kvantilregresjon (som ble omtalt i forrige kapittel).

Akkurat som for vanlig regresjonsanalyse er vanlig tidsrekkeanalyse gjennomsnittbasert. En tidsrekkemodell tilpasses observasjonene av en tidsrekke, og basert på modellen kan vi for eksempel predikere hva den neste eller de neste verdiene vil bli. Den klassiske metoden går da ut på at vi tilpasser en modell ut fra minste kvadraters metode, og regner ut forventet verdi av prediktorvariabelen, altså den verdien vi gjennomsnittlig vil få, betinget på modellen og på hittil observerte verdier. Akkurat som for regresjonsanalyse kan det i visse situasjoner, når vi har skjev fordeling, være slik at medianbasert prediksjon vil bedre fortelle oss hvordan utviklingen er sentralt i fordelingen. I tillegg vil det også her være interessant å ikke bare estimere sentrum, men hele fordelingen til prediktorvariabelen, det vil si hva de ulike kvantilene vil være. For eksempel innenfor økonomi vil man kunne få situasjoner hvor fordelingen til prediktorvariabelen vil være ikke-symmetrisk. I innledningen i [Koenker og Xiao 2006] blir det for eksempel påpekt: "*It is widely acknowledged-that many important economic variables may display asymmetric adjustment paths*",

hvorpå det blir henvist til litteratur om temaet. Og ellers i for eksempel risikoanalyse vil man kunne være interessert i hvordan ting i verste fall kan utvikle seg i den ene enden av halen, og da vil kvantilregresjonsbasert prediksjon kunne være nyttig. Har vi normalfordeling eller tilnærmet normalfordeling, vil vi som med regresjonsanalyse kunne estimere hva kvantilene vil være ved konfidensintervall, men har vi skjevhet vil denne metoden kunne gi dårlige estimater, slik at vi generelt når det gjelder dette kan gå via den kvantilregresjonsbaserte metoden.

For å ha et utgangspunkt skal vi først forklare hva tidsrekker er og beskrive noen enkle modeller samt helt grunnleggende ting. For videre å ha et referansepunkt vil vi også se hvordan man tilpasser en modell basert på tradisjonell minste kvadraters metode i det enkleste tilfellet, og hvordan man kan bruke modellen til å predikere utviklingen videre. Deretter vil vi formulere en metode basert på kvantilregresjon, slik at vi i tillegg til å erstatte gjennomsnitt med median kan få estimert hele fordelingen til prediktorvariabelen.

Stoffet som omhandler vanlig gjennomsnittbasert tidsrekkeanalyse er hentet fra [Brockwell og Davis 2002], mens stoffet som omhandler kvantilregresjonsbasert tidsrekkeanalyse[31] er hentet fra [Koenker og Xiao 2006] og [Koenker 2005], kapittel 8.3.

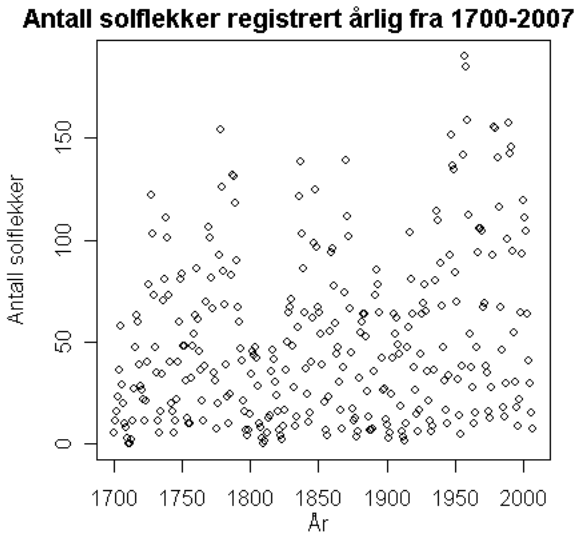
5.1 Definisjon av tidsrekke

En tidsrekke er en sekvens av observasjoner $\{x_t\}$. Indeksen, generelt gitt ved t, gir oss tidspunktet da observasjonen ble observert, altså er observasjonene sortert i tid. Vi skal kun se på tidsrekker som er diskrete i tid. Observasjonene står i sammenheng med hverandre, og det dreier seg som regel om å observere et bestemt fenomen til forskjellige tidspunkter. Ofte er vi ute etter å observere noe med jevne mellomrom, for eksempel ett år mellom hver observasjon, men det er ikke nødvendig, og modeller kan tilpasses selv om vi skulle mangle observasjoner for noen år. Vi vil her kun se på tidsrekker som har observasjoner med jevne mellomrom.

Vi kan også si at en tidsrekke er en sekvens stokastiske variabler $\{X_t\}$, og at en observert tidsrekke $\{x_t\}$ er én mulig realisering av et uendelig antall mulige observerte tidsrekker man *kunne* fått. Et viktig formål med tidsrekkeanalyse er å bruke observerte data til å på best mulig måte prøve å predikere hva som vil skje i framtiden. Det som skjer i framtiden kan vi da se på som en stokastisk variabel, men som kan være mer eller mindre avhengig av det som har skjedd før.

[31] Vi vil her ikke gå via spesialtilfellet medianen.

5.1.1 Eksempel på tidsrekke: Solflekker



***Figur 5.1**: Antall solflekker registrert hvert år i perioden 1700-2007. (kilde: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_NUMBERS/YEARLY.PLT)*

En solflekk er noe vi observerer som en flekk på sola, noe som kommer av at temperaturen i det området er lavere enn rundt. En slik flekk varer gjerne i størrelsesorden en måneds tid før den forsvinner. I datasettet om disse solflekkene som vi bruker er det 308 observasjoner, fra 1700 til 2007, som viser hvor mange solflekker det er observert hvert av årene. Se appendikset, seksjon A.2, for hele datasettet. Vi skal i seksjon 5.3 bruke denne tidsrekken som eksempel på hvordan vi tilpasser enkle modeller og predikerer framtidige verdier. Som vi ser, har vi en skjev fordeling med en del observasjoner som er større enn de fleste observasjonene. Vi skal da se hvordan dette vil spille inn når vi sammenligner den klassiske metoden med minste kvadraters metode for å tilpasse modeller, med en metode basert på kvantilregresjon. Som et alternativ til den klassiske metoden kan vi da regne ut medianestimat, og i tillegg kan vi se på hele fordelingen til prediktorvariabelen og ikke bare sentralt.

Kovarians og korrelasjon

I tidsrekkeanalyse har vi ofte bruk for å vise til sammenheng mellom de stokastiske variablene tidsrekken i utgangspunktet består av. Kovarians er et avhengighetsmål som måler lineær sammenheng mellom stokastiske variabler.

Kovariansen mellom to stokastiske variabler, for eksempel X og Y, skriver vi som $\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))]$. Av dette ser vi at $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$. Videre har vi $\mathrm{Var}(X) = \mathrm{E}[(X-\mathrm{E}(X))^2] = \mathrm{Cov}(X,X)$.

Vi har ellers $\mathrm{Var}(X) = \mathrm{Cov}(X,X) \geq \mathrm{Cov}(X,Y)$.

Vi har også $\mathrm{Cov}(X,Y) = \mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))] = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$, og

i situasjoner senere vil vi ofte se at vi har $\mathrm{E}(X) = 0$ og $\mathrm{E}(Y) = 0$, noe som da gjør at $\mathrm{Cov}(X, Y) = \mathrm{E}(XY)$.

Korrelasjon mellom to stokastiske variabler defineres som $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$, og vi får da betingelsen $|\mathrm{Corr}(X, Y)| \leq 1$. For mer om kovarians og korrelasjon, se for eksempel [Walpole m.fl 2002] side 101-102.

5.1.2 De enkleste tidsrekkemodellene

Vi vil for enkelhets skyld videre jobbe med modeller som har forventning lik 0. Disse modellene kan brukes på alle observerte tidsrekkesekvenser når man jobber med de klassiske gjennomsnittsmetodene, hvis man trekker gjennomsnittet fra hver observerte verdi. Dette får ingen konsekvenser for analysene.

Uavhengig, identisk fordelt støy, og hvit støy
Har vi en tidsrekke $\{X_t\}$ der $\mathrm{E}(X_t) = 0$ og $\mathrm{E}(X_t^2) = \sigma^2 < \infty$, samt at X_r og X_s er uavhengige for alle $r \neq s$, så har vi uavhengig, identisk fordelt støy (IID noise, independent and identically distributed noise), altså ingen sammenheng mellom hver enkelt av variablene.

Har vi i stedet for uavhengighet kravet $\mathrm{Cov}(X_r, X_s) = 0$ for alle $r \neq s$, så har vi det vi kaller hvit støy (WN, white noise). Vi ser at IID støy impliserer WN, men ikke motsatt, siden kovarians lik 0 bare forteller oss at det ikke er noen *lineær* avhengighet.

AR(1)-modellen, første ordens autoregresjon
Vi har modellen:

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \ldots ,$$

der $\{Z_t\} \sim WN(0, \sigma^2)$, $\phi < |1|$ og $\mathrm{Cov}(Z_t, X_s) = 0$ for alle $s < t$. Kravet om at $|\phi| < 1$ skyldes at om vi har $|\phi| > 1$, så vil det bety at modellen impliserer at X_t går mot pluss eller minus uendelig etter som t øker, noe som strider mot stasjonaritetsbegrepet vi innfører i neste underseksjon. Enhetsroten $|\phi| = 1$ gir heller ikke stasjonaritet, se [Brockwell og Davis 2002] side 194-196 for mer om enhetsrøtter.

Generelt har vi en AR(p)-modell, der hver observasjon er avhengig av de p foregående, pluss et støyledd. Vi skriver dette som:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t, \quad t = 0, \pm 1, \ldots ,$$

med krav om at røttene i det karakteristiske polynomet $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ ligger utenfor den komplekse enhetssirkelen[32], dvs $\phi(z) \neq 0$ for $|z| \leq 1$. Dette kravet er av samme grunn som kravet $|\phi| < 1$ for AR(1)-modellen.

[32] Se for eksempel [Brown og Churchill 2003] for en innføring i komplekse tall og kompleks funksjonsteori.

MA(1)-modellen, første ordens glidende gjennomsnitt
Vi har modellen:

$$X_t = Z_t + \theta Z_{t-1}, \quad t = 0, \pm 1, \ldots ,$$

der $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ og θ er en reell konstant.

Vi ser her at i følge modellen er hver observasjon lik en konstant multiplisert med forrige støyledd pluss et nytt støyledd. Vi kaller modellen MA(1), og på samme måte som for autoregresjon har vi generelt MA(q), som skrives som:

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}, \quad t = 0, \pm 1, \ldots .$$

ARMA(1,1)-modellen, kombinasjon av AR(1) og MA(1)
Vi har modellen:

$$X_t = \phi X_{t-1} + Z_t + \theta Z_{t-1}, \quad t = 0, \pm 1, \ldots$$

der $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $|\phi| < 1$ og $\phi + \theta \neq 0$. Vi må ha $|\phi| < 1$ for at prosessen skal være kausal, det vil si at vi skal kunne skrive prosessen som en sum av tidligere observerte verdier og ikke framtidige.

På samme måte som før har vi generelt en ARMA(p, q)-modell, som skrives som:

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \ldots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}, t = 0, \pm 1, \ldots .$$

Her har vi det grunnleggende kravet om at de karakteristiske polynomene $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ og $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ ikke skal ha noen felles løsninger. Se [Brockwell og Davis 2002] side 55-57 for flere detaljer om restriksjonene til ϕ og θ. Vi har i tillegg flere mer avanserte modeller enn dette, som vi ikke skal komme nærmere inn på her. Igjen henvises det til [Brockwell og Davis 2002].

Hva modellene brukes til
Hovedformålet ved tidsrekkeanalyse er å tilpasse en av alle de mulige modellene til en observert tidsrekke, og vi kan videre for eksempel bruke modellen til å predikere framtidsverdier. Vi må da se på hvordan vi kan finne ut hvilken modell som passer best. Det kan gjøres ved at vi ut fra observasjonene i en tidsrekke og en modell regner ut de estimerte residualene, $\{\hat{z}_t\}$, og ser om de oppfyller kravene til hvit støy. For en AR(1)-modell for eksempel, vil den t'te residualen (det vil si det t'te estimerte støyleddet) være lik $\hat{Z}_t = x_t - \hat{\phi} x_{t-1}$. Skulle det for eksempel vise seg at residualene er tydelig korrelerte, tyder det på at modellvalget er dårlig, og en ny modell bør da velges. Vi skal konsentrere oss om AR(1)-modellen, og den kvantilregresjonsbaserte

søstermodellen QAR(1). For mer om hvordan man kan sjekke hvilken modell som passer best, se [Brockwell og Davis 2002], kapitlene 1.6 og 5.5. Se også seksjonen om autokovariansfunksjonen under.

5.1.3 Stasjonaritet

Vi lar $\{X_t\}$ være en tidsrekke med $\mathrm{E}(X_t^2) < \infty$.
Vi kaller $\mu_X(t) = \mathrm{E}(X_t)$ for forventningsfunksjonen til $\{X_t\}$. For å ha stasjonaritet krever vi at $\mu_X(t)$ skal være uavhengig av t.

Kovariansfunksjonen til $\{X_t\}$ er $\gamma_X(r,s) = \mathrm{E}[(X_r - \mu_X(r))(X_S - \mu_X(s))] = \mathrm{Cov}(X_r, X_s)$, der r og s er heltall. For at tidsrekken skal være stasjonær krever vi videre at $\mathrm{Cov}(X_r, X_{r+h}) = \mathrm{Cov}(X_s, X_{s+h})$, der h er et heltall.

Vi har i tillegg et sterkere stasjonaritetsbegrep, som vi kaller streng stasjonaritet. Her kreves det at $(X_1, ..., X_n)$ og $(X_{1+h}, ..., X_{n+h})$ har samme simultanfordeling for alle h og n. Dersom vi også har $\mathrm{E}(X_t^2) < \infty$ for alle t, så vil streng stasjonaritet implisere det vi ovenfor har definert som stasjonaritet, som av og til kalles svak stasjonaritet. Vi vil videre bruke omgrepet stasjonaritet om den svake formen.

Modellene vi så på i forrige underseksjon er stasjonære. Fordelen med stasjonære modeller er at vi har en sammenheng mellom variablene i tidsrekken som ikke varierer med tiden. Når modellen stemmer godt med observerte data, gir dette gode prediksjonsmuligheter.

5.1.4 Autokovariansfunksjonen og modelltilpassing

Kovariansfunksjonen og stasjonaritet ble definert i seksjon 5.1.3. Vi definerer autokovariansfunksjonen (ACVF, autocovariance function) til en stasjonær tidsrekke $\{X_t\}$ som en funksjon av h:

$\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t)$, for alle t. Regelen $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ gir oss i tillegg at $\gamma_X(h) = \gamma_X(-h)$.

Ut fra en tidsrekkemodell kan vi regne ut hva ACVF er for de ulike verdier av h. Vi skal videre se hvordan vi kan finne empirisk ACVF når vi har en observert tidsrekke. Sammenligner vi den empiriske ACVF'en vi finner med mulige ACVF'er for tidsrekkemodeller, kan vi se hvilken modell som passer best til den observerte tidsrekken.

Autokorrelasjonsfunksjonen (ACF) blir $\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$.

Hvit støy

Hvis $\{X_t\}$ er hvit støy og $\mathrm{E}(X_t) = \sigma^2 < \infty$, har vi at $\gamma_X(h) = \left\{ \begin{array}{ll} \sigma^2, & h = 0 \\ 0, & h \neq 0 \end{array} \right\}$. Vi har dermed oppfylt kravene for stasjonaritet, siden $\mathrm{E}(X_t) = 0$ for alle t.

AR(1)-modellen[33]

Vi kan regne ut ACVF og ACV og vise at AR(1)-modellen er stasjonær, dersom $|\phi| < 1$. Fra [Brockwell og Davis 2002], side 17-18 finner vi en kort utledning som viser at $\gamma_X(h) = \frac{\phi^{|h|}\sigma^2}{1-\phi^2}$, der h er et heltall, og dette sammen med det faktum at $\mathrm{E}(X_t) = 0$ viser at AR(1)-modellen er stasjonær.

Vi får også at $\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \phi^{|h|}$, for $h = 0 \pm 1, \ldots$, et resultat vi får bruk for senere.

Den empiriske autokovariansfunksjonen

I forrige avsnitt fant vi autokovariansfunksjonen for en AR(1)-modell. Når vi skal jobbe med en observert tidsrekkesekvens, så har vi i utgangspunktet som regel ingen modell. Vi kan da gå andre veien, nemlig å beregne den empiriske autokovariansfunksjonen (dvs empirisk ACFV), og prøve å se hvilken modell som passer best ut fra resultatet vi da får, dvs sammenligne med ACVF til aktuelle modeller. Det finnes mange andre muligheter for å finne modeller, og å beregne koeffisientene i modeller, men her vil vi bare se på det helt grunnleggende. For mer informasjon om muligheter som finnes, se for eksempel [Brockwell og Davis 2002].

Empirisk ACVF[34] er:

$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \overline{x})(x_t - \overline{x})$, $-n < h < n$.

Empirisk ACV, som man kan velge å jobbe med i stedet, blir da:

$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$, $-n < h < n$.

Den partielle autokovariansfunksjonen

For å identifisere hvilken modell som passer best, kan vi også benytte oss av den partielle autokovariansfunksjonen (PACF). Vi skriver PACF som $\alpha(\cdot)$, en funksjon definert ved ligningene $\alpha(0) = 1$ og $\alpha(h) = \phi_{hh}, h \geq 1$, der ϕ_{hh} er den siste komponenten av $\phi_h \Gamma_h^{-1} \gamma_h$, der $\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$, og $\gamma_h = [\gamma(1), \gamma(2), ..., \gamma(h)]^\top$.

Empirisk PACF finner man ved å sette inn de empiriske autokovariansfunksjonene

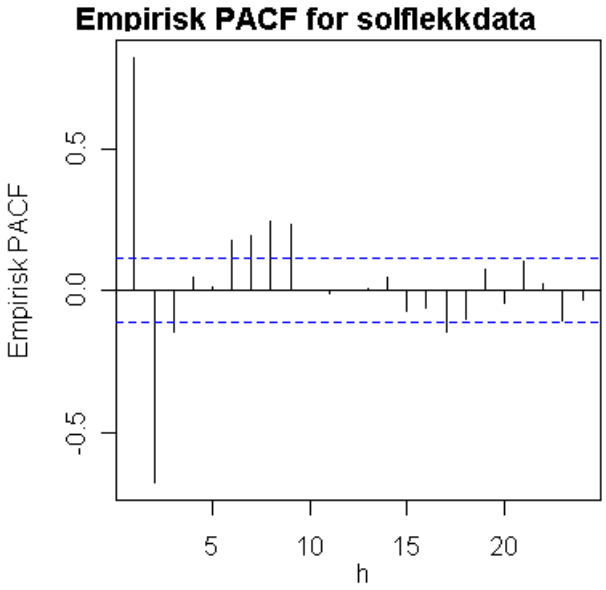
[33] Definert i detalj på side 65.

[34] Se [Brockwell og Davis 2002] sidene 19 og 59-63.

for hver av ligningene over. For en AR(p)-modell viser det seg at PACF har følgende egenskaper[35]:

$\alpha(p) = \phi_p$ og $\alpha(h) = 0$ for $h > p$.

Dette kan utnyttes til å for eksempel finne ut hvilken AR(p)-modell som passer best til en tidsrekke. Plotter man empirisk PACF vil vi kunne se hvilken p som passer. Dette har vi gjort i `R` ved funksjonen `pacf`, for solflekktidsrekken vi introduserte på side 64. Se Figur 5.2 under:



Figur 5.2: *Empirisk PACF for solflekkdataene.*

Vi har et 0.95 konfidensbånd[36] merket av på figuren, som regnes ut ved $\pm\frac{1.96}{\sqrt{n}} = \pm\frac{1.96}{\sqrt{308}} \approx \pm 0.11$. Her ser vi at vi får store utslag for $h = 1$ og $h = 2$, og mindre utslag for $h > 2$, slik at det kan se ut som at AR(2) kan fungere som en tilnærmet riktig autoregressiv modell. Vi har en del verdier for $h > 2$ som havner utenfor konfidensbåndet, og vi ser en klar tendens i verdiene vi får for $h > 2$, noe som tyder på at en ren autoregressiv modell ikke er den beste. Det viser seg at det finnes ikke-lineære modeller som passer bedre for denne tidsrekken (se for eksempel [Brockwell og Davis 2002] side 344). Vi skal ikke gå nærmere inn på slike modeller her. Slike tidsrekker blir likevel ofte brukt som eksempel når man tar for seg lineære

[35]Se [Brockwell og Davis 2002] Example 3.2.6 på side 95.
[36]Se [Brockwell og Davis 2002] Example 3.2.7 på side 96 for hvordan dette kan vises.

modeller (se for eksempel [Brockwell og Davis 2002] side 81 og 99). Grunnen til at vi har valgt denne tidsrekken å se nærmere på her, er den klare skjevheten i de observerte dataene, med noen relativt få observasjoner som er mye større enn de fleste andre, se Figur 51 på side 64.

Hvordan tilpasse modell
En måte å beregne et estimat for ϕ på i en AR(1)-modell, er å kjøre en enkel, lineær regresjonsanalyse hvor vi minimerer total kvadratavstand, altså med hensyn på $\hat{\phi}$ minimere uttrykket:

$\sum_{t=2}^{n}(x_t - \hat{\phi}x_{t-1})^2$.

Vi husker fra regresjonsanalyse at vi observerer en responsverdi, samt tilhørende forklaringsvariable, der vi antar at hver observasjon er uavhengig av de andre. I tidsrekkeanalyse observerer vi en verdi, men her er det altså *andre observasjoner,* for en AR(1)-modell den foregående observasjonen, som kan sies å være forklaringsvariable, slik at det er en sterk avhengighet innad i modellen. Skal vi foreta regresjonsanalyse på et datasett, når vi for eksempel har bestemt oss for å bruke en AR(1)-modell, så vil det i praksis si at vi må se på de $n-1$ observasjonsparene[37] vi har. Dette blir i stedet for, når vi har regresjonsanalyse med én forklaringsvariabel, å se på de n parene med responsverdier og tilhørende verdier av forklaringsvariabelen. Vi skal her altså komme fram til et estimat for ϕ. Plotter vi alle disse observasjonsparene (x_{t-1}, x_t), så vil en minste kvadraters regresjonslinje gi et estimat for stigningstallet i grafen, som i praksis vil være det vi må gange x_{t-1} med for å få et estimat av x_t, altså $\hat{\phi}$.

Etter at vi har tilpasset en modell kan vi sjekke om vi har fått all avhengighet mellom observasjonene med i modellen, ved å se om residualene, $\hat{x}_t = x_t - \hat{\phi}x_{t-1}$, er uavhengige av hverandre. Eventuelt kan man plotte ACF som man får fra modellen og sammenligne med empirisk ACF, noe som er vist i [Brockwell og Davis 2002] på side 63.

5.1.5 Prediksjon

Når vi, for en AR(1)-modell, har funnet et estimat for ϕ-koeffisienten, så er det enkelt å foreta prediksjon. Har vi n observasjoner i tidsrekkesekvensen, så vil vi få $\widehat{\mathrm{E}(x_{n+1}|x_n)} = \hat{\phi}x_n$, siden modellen antar $\mathrm{E}(Z_t) = 0$. Vi husker fra seksjon 2.1.2 at ved å minimere kvadratavstand får vi forventningsverdi. Skal vi predikere m skritt framover i tid, vil vi få $\widehat{\mathrm{E}(x_{n+m}|x_n)} = \hat{\phi}^m x_n$.

[37] Det er dette som er grunnen til summen over går fra 2 til n.

5.2 Kvantil autoregresjon

I seksjon 5.1.4 så vi at ved minste kvadraters regresjonsanalyse kunne vi finne et estimat for ϕ-koeffisienten i en AR(1)-modell, og at vi ved å bruke denne koeffisienten kunne estimere forventningen[38] til framtidige verdier av tidsrekken. Igjen er det to motivasjonsfaktorer for å innføre en kvantilbasert metode. For det første kan det være ønskelig å estimere hele fordelingen til prediktorvariabelen i stedet for bare forventningsverdien. I mange situasjoner kan det hende vi vil være interessert i å vite noe om hva som i verste/beste fall vil skje, og det vil ikke en estimert forventningsverdi si så mye om. Og for det andre, i situasjoner hvor fordelingen ikke er symmetrisk, så har vi tidligere argumentert for at et medianestimat bedre kan vise sentralitet enn en estimert forventningsverdi.

5.2.1 Hva har vært gjort før?

For ordinær regresjonsanalyse forutsetter man uavhengig, identisk normalfordelte støyledd. For ordinær tidsrekkeanalyse forutsetter man også uavhengig, identisk fordelte støyledd (eventuelt ukorrelerte). I begge situasjoner har man utviklet metoder som passer bedre for skjeve fordelinger, hvor man ikke forutsetter normalitet, men beholder forutsetningen om identisk, uavhengig fordelte støyledd (eller hvit støy for tidsrekker). Fra innledningen av [Koenker og Xiao 2006], vedrørende kvantil autoregresjon, siterer vi: "*Curiously, however, all of the theoretical work dealing with this model (that we are aware of) focuses exclusively on the iid innovation case that restricts the autoregressive coefficients to be independent of the specified quantiles.*" Denne forutsetningen kan gi unøyaktige resultat hvis fordelingen reelt sett er skjev i forhold til antatt og hvis ϕ varierer med kvantilen. I kvantil autoregresjon, som vi nå skal se på, forutsetter man ingen spesiell fordeling for støyleddene, på samme måte som i kvantilregresjon. Men som i kvantilregresjon vil vi for hver verdi av τ forutsette at vi har støyledd hvis τ-kvantil er 0. Altså, som i kvantilregresjon definerer vi egne støyledd for hver τ-verdi.

5.2.2 QAR(1)-modellen, kvantil autoregresjon og prediksjon

Vi ser her på den enkleste modellen, QAR(1)-modellen, som er et kvantilregresjonsbasert alternativ til tilsvarende AR(1)-modell. Selve modellen er den samme, bortsett fra at vi ikke antar noen spesiell fordeling for støyleddene, men derimot som nevnt at for hver verdi av τ forutsetter vi støyledd hvis τ-kvantil er lik 0. Vi skriver her generelt hvordan kvantilen vil se ut:

$$Q_{x_t}(\tau|x_{t-1}) = \phi_0^{(\tau)} + \phi_1^{(\tau)} x_{t-1}, \;\; t = 0, \pm 1, \ldots \; .$$

[38] Forventning basert på den observerte tidsrekkesekvensen og valg av modell.

Vi ser at vi har med konstantleddet $\phi_0^{(\tau)}$, noe som her er nødvendig. For minste kvadraters baserte modeller kunne vi sløyfe dette ved å trekke gjennomsnittet fra hver observasjon, noe som ikke vil spille inn dersom vi senere skal tilpasse modell ved minste kvadraters metode, som nettopp er en gjennomsnittsmetode. I kvantilregresjon ville vi da måttet tvinge hver regresjonslinje ikke bare gjennom origo, men også gjennom to datapunkt, og det er ikke sikkert man i det hele tatt finner punkter som ligger slik på en linje. I eksempelet vist på Figur 4.11 på side 58 hadde vi en tilsvarende situasjon, hvor kvantilregresjonslinjene måtte gjennom origo samt to datapunkter, men dette ble da ikke noe problem siden vi i den situasjonen hadde flere datapunkter som lå nettopp i origo, med både inntekt og formue lik 0. Noe lignende skjer ikke for solflekkdataene når vi trekker gjennomsnittet fra hver observasjon, og er ellers ytterst sjeldent for tidsrekker, siden vi (for en QAR(1)-modell) må få en sekvens av to observasjoner som begge må være 0 etter at gjennomsnittet er trukket fra for at vi skal få et datapunkt i origo, når vi plotter x_{t-1} mot x_t.

Her antar vi altså at begge ϕ-parametrene kan variere med τ. Vi regner ut kvantilestimater for $\phi_0^{(\tau)}$ og $\phi_1^{(\tau)}$ ved å minimere total vektet absoluttavstand med hensyn på $\hat{\phi}_0^{(\tau)}$ og $\hat{\phi}_1^{(\tau)}$:

$$(1-\tau) \sum_{x_t < \hat{\phi}_0^{(\tau)} + \hat{\phi}_1^{(\tau)} x_{t-1}} |x_t - \hat{\phi}_0^{(\tau)} - \hat{\phi}_1^{(\tau)} x_{t-1}| + \tau \sum_{x_t > \hat{\phi}_0^{(\tau)} + \hat{\phi}_1^{(\tau)} x_{t-1}} |x_t - \hat{\phi}_0^{(\tau)} - \hat{\phi}_1^{(\tau)} x_{t-1}|$$

Dette gjøres som for kvantilregresjon, også i `R`. Skal man predikere, gjøres dette på samme måte som vi så for forventningsverdier i seksjon 5.1.5. Det vil bli noe annerledes siden vi har et konstantledd, men i prinsippet gjøres det på samme måte. For prediksjon ett steg fram i tid får vi $\widehat{Q_{x_{n+1}}}(\tau|x_n) = \hat{\phi}_0^{(\tau)} + \hat{\phi}_1^{(\tau)} x_n$. Videre for prediksjon to steg fram i tid får vi $\widehat{Q_{x_{n+2}}}(\tau|x_n) = \hat{\phi}_0^{(\tau)} + \hat{\phi}_1^{(\tau)}(\hat{\phi}_0^{(\tau)} + \hat{\phi}_1^{(\tau)} x_n)$, og videre $\widehat{Q_{x_{n+m}}}(\tau|x_n) = \hat{\phi}_0^{(\tau)}(1 + \hat{\phi}_1^{(\tau)} + (\hat{\phi}_1^{(\tau)})^2 + ... + (\hat{\phi}_1^{(\tau)})^{m-1}) + (\hat{\phi}_1^{(\tau)})^m x_n$ m steg fram i tid.

5.2.3 Problemer med kvantil autoregresjon

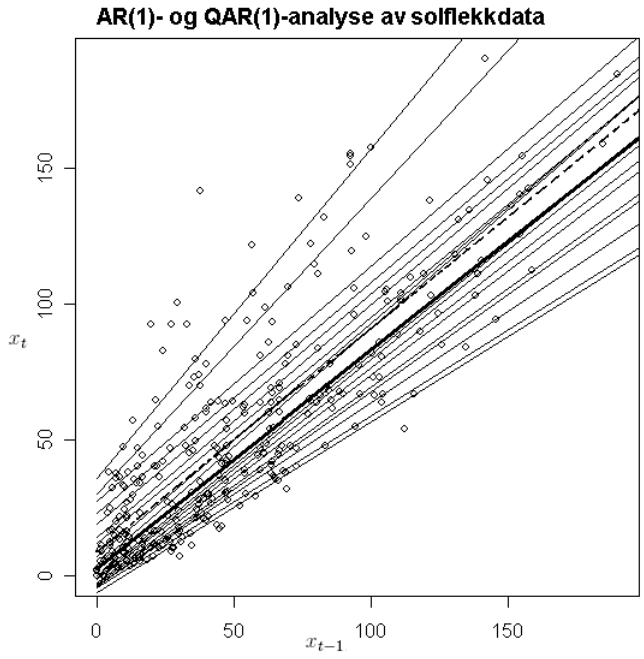
Etter som kvantilregresjon ble introdusert i 1978, og i og med at det virker temmelig rett fram å skulle overføre teorien til bruk i tidsrekkeanalyse, så kan vi spørre oss hvorfor dette ikke har skjedd tidligere. Roger Koenker har stått for introduksjonen av begge deler, med et tidssprang på nærmere 30 år. Hvorfor så lang tid? Har de eksisterende metodene innenfor tidsrekkeanalyse, eller andre metoder som har vært utviklet i løpet av denne tiden, vært gode nok? Er nyvinningen med kvantil autoregresjon begrenset, sammenlignet med hva tilfellet var for kvantilregresjon versus minste kvadraters regresjon? Dukker det opp problemer når dette overføres til tidsrekkeanalyse?

Monotonitetsbetingelse

Vi har en monotonitetsbetingelse, som sier at for relevante verdier av x_{t-1}, så må vi ha $\phi^{(\tau_1)}x_{t-1} > \phi^{(\tau_2)}x_{t-1}$ for $\tau_1 > \tau_2$, og dette er for at de ulike kvantilestimatene ikke skal krysse hverandre i kjerneområdet. I områder hvor betingelsen ikke er oppfylt, noe som kan skje i utkanten, så vil ting kunne gå galt. Vi skal se på hvorfor denne betingelsen er spesielt viktig for kvantil autoregresjon, sammenlignet med for kvantilregresjon, se under drøftingen av eksempelet i seksjon 5.3 og konklusjonen i seksjon 5.4.

En hovedforskjell på kvantilregresjon og kvantil autoregresjon er som vi har nevnt at i sistnevnte tilfelle er det mye større avhengighet i selve modellen. I regresjonsanalyse er hver observasjon med tilhørende forklaringsvariable uavhengige av de andre observasjonene. I tidsrekkeanalyse er det som regel én eller flere av de andre observasjonene som er forklaringsvariable for hver observasjon. I neste seksjon skal vi i forbindelse med et eksempel se på hvilke konsekvenser dette kan få, i sammenheng med monotonitetsbetingelsen, for den kvantilregresjonsbaserte tidsrekkeanalysen.

5.3 Eksempel på datasett



Figur 5.3: *Vi har her x_{t-1} plottet mot x_t, for $t = 2, .., 308$, og kvantilregresjon samt minste kvadraters regresjon utført på dataene. I tillegg til minste kvadraters regresjonslinjen (stiplet linje) er 19 kvantilregresjonslinjer tegnet inn ($\tau = 0.05, \tau = 0.10, ..., \tau = 0.95$), medianlinjen er markert med fet linje.*

Vi tar for oss tidsrekken om årlig registrering av antall solflekker i perioden 1700-2007, som vist på Figur 5.1 på side 64, og som er listet opp i appendikset i seksjon A.2. Som nevnt på side 69 viser det seg at ikke-lineære modeller gir de aller beste resultatene for denne tidsrekken, men slike tidsrekker blir likevel ofte brukt som eksempel når man tar for seg lineære modeller. Vi kaller tidsrekken for $\{x_t\}$, der x_1 er antall solflekker observert i 1700, og så videre til x_{308} som er tilsvarende for 2008.

Vi skal tilpasse en AR(1)-modell til $\{x_t\}$, og deretter tilpasse en QAR(1)-modell for $\tau = 0.05, \tau = 0.10, ..., \tau = 0.95$. Vi bruker `R`, og benytter oss av henholdsvis minste kvadraters regresjons- og kvantilregresjonsrutinene for å finne estimater, slik vi har beskrevet tidligere. Vi jobber direkte med tidsrekken vi har selv om vi ikke har forventning lik 0, og vi må derfor ha med et konstantledd. Kaller vi forventningen μ, får vi da AR(1)-modellen $X_t - \mu = \phi_1(X_{t-1} - \mu) + Z_t$, og setter vi $\phi_0 = \mu(1 - \phi_1)$ får vi $X_t = \phi_0 + \phi_1 X_{t-1} + Z_t$, som vi vil bruke i utregningene. Tilsvarende får vi $Q_{x_t}(\tau|x_{t-1}) = \phi_0^{(\tau)} + \phi_1^{(\tau)} x_{t-1}$, som altså er τ-kvantilen i QAR(1)-modellen, som vist i forrige seksjon. Regresjonslinjene vi får er vist i et plott i Figur 5.3 på forrige side, med alle observasjonsparene (x_{t-1}, x_t), hvor altså stigningstallet til linjene tilsvarer $\hat{\phi}_1$-koeffisienten, og skjæringspunktet med x_t-aksen tilsvarer $\hat{\phi}_0$-koeffisienten. Tolkningen av en linje er at når vi har gitt x_{t-1}, så viser linjen et estimat av hva x_t vil bli.

Vi merker oss at selv om vi ser av Figur 5.1 på side 64 at vi har en svært skjev fordeling hvor flesteparten av observasjonene ligger nokså lavt, så ser vi av Figur 5.3 på forrige side at når vi plotter x_{t-1} mot x_t, så blir det nokså symmetrisk. Grunnen til at det blir slik, er at om man har et observasjonspar (x_{t-1}, x_t), hvor for eksempel x_t er en uteliggende observasjon, vil vi i neste observasjonspar, (x_t, x_{t+1}), hvis x_{t+1} ikke er en uteligger, få et observasjonspar som mer eller mindre kan nøytralisere det forrige observasjonsparet. Det vil si at disse to observasjonsparene da vil ligge på hver sin side av sentrum, og når vi plotter x_{t-1} mot x_t får vi to uteliggere som ligger noenlunde like langt ute på hver side, og minste kvadraters metoden vil da ikke være så sårbar. Naturligvis avhenger dette av at de tilhørende observasjonene, altså x_{t-1} og x_{t+1}, ligger omtrent i samme område. Dette trenger de ikke å gjøre, så mulighet for skjevhet er fremdeles til stede. Like fullt vil vi hevde avhengigheten mellom observasjoner som ligger i en tidsrekke begrenser muligheten for ikke-symmetri hos prediktorvariablene. Dette kan være en grunn til at kvantilregresjonsmetoden kan ha begrenset nytteverdi for tidsrekkeanalyse, nemlig at det skal mer til for å få skjevhet på grunn av denne avhengigheten mellom observasjonspar.

Ett-stegspredikorvariabelen

Når vi bruker QAR(1)-modellen til prediksjon, får vi ganske enkelt et τ-estimat av ett-stegsprediktorvariabelen ved å regne ut $\widehat{Q_{x_{309}}}(\tau|x_{308}) = \hat{\phi}_1^{(\tau)} + \hat{\phi}_1^{(\tau)} x_{308}$, der vi har $x_{308} = 7.5$, som er siste observerte verdi. Dette er en relativt sett svært lav verdi. Vi har foreløpig bare forholdt oss til AR(1)-modellen, men som vi så under teorien om PACF på side 68, så viser det seg at en AR(2)-modell er bedre enn AR(1)-modellen

for solflekkdataene. I Tabell 5.1 under tas derfor resultater for AR(2)-modellen også med, hvor verdiene vi får også vil være betinget på $x_{307} = 15.2$, i tillegg til x_{308}.

AR(1)	AR(2)
$\widehat{Q_{x_{309}}}(\tau = 0.05 \mid x_{308}) = -1.848164$	$\widehat{Q_{x_{309}}}(\tau = 0.05 \mid (x_{307}, x_{308})) = -1.932692$
$\widehat{Q_{x_{309}}}(\tau = 0.10 \mid x_{308}) = -0.1920273$	$\widehat{Q_{x_{309}}}(\tau = 0.10 \mid (x_{307}, x_{308})) = 1.405181$
$\widehat{Q_{x_{309}}}(\tau = 0.15 \mid x_{308}) = 0.8027671$	$\widehat{Q_{x_{309}}}(\tau = 0.15 \mid (x_{307}, x_{308})) = 2.505$
$\widehat{Q_{x_{309}}}(\tau = 0.20 \mid x_{308}) = 0.906422$	$\widehat{Q_{x_{309}}}(\tau = 0.20 \mid (x_{307}, x_{308})) = 3.889639$
$\widehat{Q_{x_{309}}}(\tau = 0.25 \mid x_{308}) = 1.699602$	$\widehat{Q_{x_{309}}}(\tau = 0.25 \mid (x_{307}, x_{308})) = 5.332243$
$\widehat{Q_{x_{309}}}(\tau = 0.30 \mid x_{308}) = 2.6714$	$\widehat{Q_{x_{309}}}(\tau = 0.30 \mid (x_{307}, x_{308})) = 5.687948$
$\widehat{Q_{x_{309}}}(\tau = 0.35 \mid x_{308}) = 4.049765$	$\widehat{Q_{x_{309}}}(\tau = 0.35 \mid (x_{307}, x_{308})) = 7.322224$
$\widehat{Q_{x_{309}}}(\tau = 0.40 \mid x_{308}) = 4.649362$	$\widehat{Q_{x_{309}}}(\tau = 0.40 \mid (x_{307}, x_{308})) = 8.425684$
$\widehat{Q_{x_{309}}}(\tau = 0.45 \mid x_{308}) = 5.707902$	$\widehat{Q_{x_{309}}}(\tau = 0.45 \mid (x_{307}, x_{308})) = 9.624742$
$\widehat{Q_{x_{309}}}(\tau = 0.50 \mid x_{308}) = 8.173862$	$\widehat{Q_{x_{309}}}(\tau = 0.50 \mid (x_{307}, x_{308})) = 11.25466$
$\widehat{Q_{x_{309}}}(\tau = 0.55 \mid x_{308}) = 9.834635$	$\widehat{Q_{x_{309}}}(\tau = 0.55 \mid (x_{307}, x_{308})) = 13.45468$
$\widehat{Q_{x_{309}}}(\tau = 0.60 \mid x_{308}) = 12.72365$	$\widehat{Q_{x_{309}}}(\tau = 0.60 \mid (x_{307}, x_{308})) = 14.97419$
$\widehat{Q_{x_{309}}}(\tau = 0.65 \mid x_{308}) = 17.42206$	$\widehat{Q_{x_{309}}}(\tau = 0.65 \mid (x_{307}, x_{308})) = 16.66425$
$\widehat{Q_{x_{309}}}(\tau = 0.70 \mid x_{308}) = 20.12951$	$\widehat{Q_{x_{309}}}(\tau = 0.70 \mid (x_{307}, x_{308})) = 18.46887$
$\widehat{Q_{x_{309}}}(\tau = 0.75 \mid x_{308}) = 24.70878$	$\widehat{Q_{x_{309}}}(\tau = 0.75 \mid (x_{307}, x_{308})) = 20.02483$
$\widehat{Q_{x_{309}}}(\tau = 0.80 \mid x_{308}) = 28.95495$	$\widehat{Q_{x_{309}}}(\tau = 0.80 \mid (x_{307}, x_{308})) = 21.81577$
$\widehat{Q_{x_{309}}}(\tau = 0.85 \mid x_{308}) = 33.16667$	$\widehat{Q_{x_{309}}}(\tau = 0.85 \mid (x_{307}, x_{308})) = 26.08049$
$\widehat{Q_{x_{309}}}(\tau = 0.90 \mid x_{308}) = 37.7$	$\widehat{Q_{x_{309}}}(\tau = 0.90 \mid (x_{307}, x_{308})) = 31.17015$
$\widehat{Q_{x_{309}}}(\tau = 0.95 \mid x_{308}) = 44.3351$	$\widehat{Q_{x_{309}}}(\tau = 0.95 \mid (x_{307}, x_{308})) = 39.5506$
$\widehat{\mathrm{E}(x_{309})} = 14.63694$	$\widehat{\mathrm{E}(x_{309})} = 14.93614$

Tabell 5.1: *Ett-stegsprediktorvariabelens kvantiler for solflekktidsrekken for både AR(1)- og AR(2)-modellen.*

Vi ser at vi får en klar høyreskjev fordeling for AR(1)-modellen. Dette kan vi også se direkte av Figur 5.3, siden vi befinner oss helt til venstre hvor $x_{t-1} = 7.5$. Hadde siste observerte verdi vært rundt midten på x_{t-1}-aksen ville vi fått en større grad av symmetri. For AR(2)-modellen blir det også høyreskjevhet, men ikke i like stor grad. Ser vi på sentralestimatene ser vi at vi får noe lavere verdi for $\hat{Q}^{(0.50)}$ enn for $\hat{\mathrm{E}}$. At det blir høyreskjevhet overrasker ikke, da vi ikke kan observere et negativt antall solflekker, og vi har i tidsrekken mange observasjoner mellom 0 og 10 hvor den påfølgende observasjonen er fem ganger så stor eller mer. Vi ser i tillegg at vi for de to laveste kvantilverdiene får negative estimater, noe som selvsagt ikke er mulig i praksis, og som er et eksempel på en svakhet ved resultater vi får helt i utkanten.

5.4 Konklusjon

Vi har monotonitetsbetingelsen, nevnt på side 73. Hvis siste observerte verdi av en tidsrekke er innenfor monotonitetsområdet, så oppstår det ingen fundamentale problemer når vi skal predikere framtidige verdier. Men i de situasjoner den siste observerte verdi er relativt stor eller liten, slik at vi befinner oss i et område hvor monotonitetsbetingelsen ikke er oppfylt, så vil resultatene vi får ved å for eksempel predikere bli høyst merkelige og inkonsistente, siden vi da ikke vil få monotone resultater i τ. For flerstegsprediksjon vil det eskalere ytterligere. Et slikt aspekt finner vi ikke i regresjonsanalyse. Der er det slik at om vi kommer ut i utkanten, hvor vi fra før av ikke har noen observasjoner, eller bare for eksempel en uteligger, så kan det forekomme krysninger. Dette betyr ikke at hele analysen vil være bortkastet, bare at resultatene godt utenfor kjerneområdet kan bli ubrukelige. For tidsrekker, derimot, er vi først i den situasjonen hvor siste observerte verdi er utenfor kjerneområdet slik at vi ikke har monotonitet, så vil hele prediksjonsanalysen kunne feile. De gamle metodene som går på forventningsverdier og kvantiler ut fra det vil da være tryggere og bedre. Og når vi i tillegg ser hvordan tilsynelatende skjevt fordelte tidsrekker ofte ikke blir så skjeve likevel når vi plotter x_{t-1} mot x_t, så spørs det om ikke det er gode grunner til at innføringen av metodene for kvantilregresjon har latt vente på seg i tidsrekkeanalyse. Fra [Koenker og Xiao 2006], seksjon 4, siterer vi:

"As in other linear quantile regression applications, linear QAR models should be cautiously interpreted as useful local approximations to more complex nonlinear global models. If we take the linear form of the model too literally, then obviously at some point (or points) there will be "crossings" of the conditional quantile functions[..]. This crossing problem appears to be more acute in the autoregressive case than in ordinary regression applications, because the support of the design space (i.e., the set of x_t's that occur with positive probability) is determined within the model."

Fra diskusjonen som følger artikkelen kommer det fram at *noe* må gjøres med den inkonsistensen som oppstår i områdene hvor kvantilene krysser, men *hva* som skal gjøres er man ikke sikker på om man har funnet et godt nok svar på ennå. Det kan jo bli sett på som et nederlag for denne nye teorien om man for eksempel sier at i de områdene hvor kvantil autoregresjon gir inkonsistens, så kan man bare bruke eksisterende metoder i stedet. Fra [Koenker og Xiao 2006], i tilsvaret til artikkelforfatterne som følger diskusjonen av den opprinnelige artikkelen, siterer vi:

"QAR models allow us to explore some features of time series that are inaccessible through classical methods, but they suffer from their own limitations, many of which have been brought out in this fruitful discussion. Much room is left for improvement; we look forward to the continuing discussion."

Det blir interessant å se hva forskerne kommer fram til i tiden som kommer.

A Appendiks

A.1 Skatteliste-datasett

Skatteliste fra inntektsåret 2006, 474 personer med postnummer 5947, menn og kvinner hver for seg, sortert etter inntekt. *(kilde: dagbladet.no/skatt/)*

					Menn						
Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt
2868832	718192	1363666	369255	185079	133295	210170	584661	56192	110466	0	45009
1324565	1302349	464580	360705	7637	131050	209355	0	69457	110358	615467	16737
1222297	2295158	380862	348847	0	130819	208881	890717	58038	109907	787903	29231
1183831	734585	510670	338455	0	132518	203879	349685	52839	105607	353386	9927
993426	337349	423064	330802	333125	115829	202195	1397119	60576	104356	20946	10458
816007	618329	344839	330458	116435	131268	202133	17137	67395	102787	0	27677
724836	0	300019	329562	573753	94063	201742	1584689	61334	101846	526614	1139
716042	695872	285811	328864	12283	107746	201133	0	60888	101397	552506	10456
673830	2613038	274982	326108	0	134734	199229	0	67079	98181	0	29432
673759	0	272344	320863	566810	86319	198974	532912	49794	97312	181670	5756
662893	1459680	271958	316032	0	97100	198583	591197	52108	96960	313238	3197
640109	245121	235740	314219	0	117253	195033	291796	46422	93213	0	27933
639746	231390	252154	313981	421322	109072	194193	0	68256	91551	428876	2528
614239	64524	232919	311038	550221	108786	192459	596289	49458	91148	1399309	31274
601004	36967	240987	309028	0	122184	191514	0	64807	90379	590696	14866
600005	0	251221	298064	0	110075	191125	276288	45969	88844	0	16158
584392	0	254593	295202	783567	82660	189028	440029	50121	87132	169971	2354
573105	105664	235859	288814	1452634	84917	187279	0	46913	85398	127689	0
565703	746098	224089	284426	188852	106292	186653	893640	52029	83238	0	23483
564116	0	218978	283405	1175812	82538	185551	0	67393	79499	15534	21302
562001	646767	211315	282595	1599793	107153	183831	0	67393	77040	0	25279
544146	2276152	173150	279767	0	104364	183296	506538	48381	76348	25281	21040
543198	0	213123	279729	105646	95870	181491	0	63059	75886	0	28200
521734	471273	137422	276960	2184698	90493	181449	710228	53206	75378	0	10552
516307	549861	201743	274998	5783097	145798	179750	0	56292	72640	15662	14637
511668	584151	199375	271276	1485158	81890	171007	0	43327	67958	0	17074
508292	2275719	206314	267992	555404	94024	170485	13513	39027	67729	0	17024
490118	0	203908	267762	0	94974	170368	651818	44271	63953	67517	0
486577	570985	184439	264905	0	92762	169082	1757884	53830	61208	172195	15806

Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt
479658	0	202046	264297	0	101379	168080	0	59870	48158	0	9871
475706	287948	161095	264108	0	94645	165052	970112	47197	46944	146877	0
472131	967790	183933	263861	1261592	97441	163471	599187	35673	45462	8533	8856
472004	0	183034	263366	616609	94118	156329	977420	39143	43441	0	8217
468172	762758	177058	262991	2485113	90695	152875	379384	34849	32883	0	4807
467975	0	199400	261926	313600	91607	151065	0	34959	24107	0	4289
453393	78272	169162	261140	0	94690	150501	0	50303	13167	10879	5321
452799	1276556	174179	259683	0	96182	146498	1572514	44270	12057	21471	375
446820	976676	171481	258867	0	90877	146478	352444	31170	10840	227015	243
445137	731168	168504	256293	828269	77029	144508	36461	27064	10558	0	6020
438572	224994	165189	256084	0	88694	143066	0	26271	3939	166656	0
438337	700528	182498	254373	171142	87743	143010	389836	34370	3417	249381	4355
435652	2257072	179287	253513	0	104190	142880	441990	31755	3059	391331	1722
430521	1745954	185571	252970	0	107292	140870	111853	25064	1044	0	0
426776	0	165008	252697	0	87880	137832	11932	48831	763	0	0
425453	189866	155945	248118	0	84266	132390	464951	25700	468	0	0
425055	0	169065	246033	1707457	77985	130741	406803	30372	415	0	0
424432	0	171327	244156	0	76798	127811	1271002	49771	160	12616	0
418366	1519160	140803	235528	0	81551	127074	346072	19057	79	12676	19
416905	414714	169962	234674	0	87183	126229	370455	20415	2	1092	0
414750	0	155635	233165	0	80062	125870	96653	24120	1	0	0
410459	1443290	116721	230797	0	50135	124112	0	33329	0	0	0
409947	362731	154850	225941	745955	63630	123631	0	29068	0	27786	0
406444	427430	147662	223606	62776	76112	120480	213011	14131	0	3594	25
401049	1781197	159482	221996	0	77495	120310	552358	20802	0	3594	25
400331	1449142	157461	221961	0	71559	119043	0	41406	0	0	0
389690	703991	160214	220572	324690	55590	118539	85164	12781	0	19061	0
378818	308703	136036	216997	0	77373	118052	108271	34056	0	189	0
378219	0	143306	213090	34756	71337	115048	0	37518	0	0	19634
372180	0	136344	212107	0	65189	113331	408184	24328	-	-	-
371120	0	139583	210977	0	76004	113288	6759	15143	-	-	-
370356	37359	136655	210823	655632	57231	111218	123775	8755	-	-	-

Tabell A.1: *Skatteliste menn, 241 personer.*

					Kvinner						
Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt
576154	9037636	277940	163114	615680	46697	103731	336217	19718	78825	173377	0
470382	0	161495	160106	82584	53650	103459	275649	5993	78604	0	0
427254	0	160195	159501	86000	34467	101145	60802	3215	78348	171500	12638
421474	0	166556	159393	193063	53811	101092	28185	30725	78000	0	21954
421343	918224	156186	159325	136160	35214	100824	145667	3038	77184	0	20976
404920	0	166	158851	0	53482	98194	106210	29627	76122	51014	1691
371292	367491	104645	157207	456738	40279	97744	535090	31519	76033	202672	13250
365130	342873	131575	156832	466430	36954	97658	93772	1297	75496	61106	11740
343398	1025532	100879	151991	0	50568	97180	203425	13726	75455	8587	452
338584	0	132093	151795	0	50939	97160	0	1023	75147	24134	11458
328131	0	116347	148616	0	48239	97023	0	15524	75142	5056	7847
325424	276883	112214	146657	253992	47299	96731	0	30967	74726	0	5861
321949	20358	115215	145639	207378	32122	96167	391088	4287	73153	63181	2530
321760	826192	116052	145028	119266	47481	95942	947796	16149	72907	0	20674
319795	101429	1103598	143978	0	26773	94656	0	20711	71917	0	8258
310065	540542	108886	143835	0	47335	94073	410506	3530	70532	0	18997
308601	426067	107552	142838	181325	37275	93925	167534	28900	67690	0	24671
308456	0	103443	140985	903781	36224	93620	395177	2977	66680	333432	17634
301208	207858	113031	139551	0	36787	93534	0	19850	62072	0	11900
299666	0	102873	138072	514231	43580	92826	0	28058	58653	0	15446
296057	0	102930	137752	0	29495	91776	506	0	56757	22	14353
295967	0	93055	137149	1052458	35421	91010	100027	26599	54623	41611	9129
295433	350138	99534	134458	0	46082	90364	25594	500	54121	210457	13087
276428	0	95375	133748	0	45963	90356	926535	31468	53018	0	11546
267041	460649	92815	133394	230481	38023	90341	8737	0	50235	72111	10591
263018	0	92728	131107	205628	42819	89680	16156	16152	47974	0	11406
262984	0	92794	130032	442875	44047	89312	38292	26550	47664	0	6725
243186	407105	84102	129867	361827	22229	88879	369229	1523	44793	0	14722
240661	0	83445	129515	322306	42803	88787	547003	15064	44058	0	5864
239084	1056824	78030	127888	30297	26671	88468	311335	1002	43377	0	12760
227512	0	76298	127293	18174	41716	88364	22388	12876	39302	0	6690
226100	672087	61131	127251	0	39671	88105	0	22921	37936	0	6209
225878	204881	76676	126165	0	40721	88043	0	25656	32189	0	5034
220772	95629	69515	122925	0	39405	88013	140998	19632	25814	0	0
215891	9910	72509	119871	170820	13514	87896	569859	3388	25030	2300409	20024
215489	911821	59885	119288	3255	38012	87135	636666	20509	21783	384575	3813

Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt	Inntekt	Formue	Skatt
213205	35577	66747	119123	41325	13103	87111	38470	26308	21732	0	0
212128	0	72260	118293	0	12646	86183	0	27397	18345	0	0
207654	379378	72408	116628	215018	25066	85152	362378	1462	17155	0	3850
202311	1016630	56525	116294	23660	37064	85052	30293	0	14613	0	3571
197695	507823	50186	114568	563788	17933	84994	50761	0	14530	618884	3928
188581	2584	62923	114453	0	37827	84928	129037	0	6978	345205	4618
188454	192953	63489	114244	410067	14619	84924	182133	0	5931	87887	1039
185391	0	61102	113179	0	37690	84856	25829	0	5460	0	1903
183697	2362	68877	112784	62582	9616	84764	108779	0	1516	0	100
180749	465903	63535	112704	476738	25499	84738	70924	24356	1185	0	830
174798	744329	64027	112084	93684	9231	84730	316651	9177	1182	91004	38
174624	1252066	50430	110323	75734	35038	84715	455580	16018	673	0	722
173898	6773	57646	109353	521520	22980	84138	16905	0	363	30495	0
173161	757426	45010	108749	425360	11901	83652	0	23134	24	3888	0
171709	359869	58655	108283	474042	23686	83625	425830	2033	8	0	2
171144	31154	56572	107554	189410	35513	83482	629772	13679	0	0	0
170856	48031	56447	107443	1229055	28648	82378	232165	289	0	245	0
168829	267771	38989	107437	1114835	26123	81707	0	10887	0	0	12326
167807	251785	45808	107328	321260	9027	81064	0	24549	0	0	0
167776	0	55754	107201	0	34742	79717	292301	22506	0	0	7932
165009	199314	37311	106355	394823	29261	79454	0	24509	0	0	0
164183	365379	56211	105349	0	36250	79118	101504	22987	-	-	-
163472	124617	36881	105108	0	18782	78864	187959	13238	-	-	-

Tabell A.2: *Skatteliste kvinner, 234 personer.*

A.2 Solflekkdata

Antall solflekker registrert årlig i perioden 1700-2008. Som vi ser har vi desimaltall fra og med 1749. Dette skyldes at man har vektet resultatene fra flere observatører. *(kilde: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_NUMBERS/YEARLY.PLT)*

År	Antall	År	Antall	År	Antall	År	Antall	År	Antall	År	Antall	År	Antall	År	Antall
1700	5	1739	101	1778	154.4	1817	41.1	1856	4.3	1895	64.0	1934	8.7	1973	38.0
1701	11	1740	73	1779	125.9	1818	30.1	1857	22.7	1896	41.8	1935	36.1	1974	34.5
1702	16	1741	40	1780	84.8	1819	23.9	1858	54.8	1897	26.2	1936	79.7	1975	15.5

1703	23	1742	20	1781	68.1	1820	15.6	1859	93.8	1898	26.7	1937	114.4	1976	12.6
1704	36	1743	16	1782	38.5	1821	6.6	1860	95.8	1899	12.1	1938	109.6	1977	27.5
1705	58	1744	5	1783	22.8	1822	4.0	1861	77.2	1900	9.5	1939	88.8	1978	92.5
1706	29	1745	11	1784	10.2	1823	1.8	1862	59.1	1901	2.7	1940	67.8	1979	155.4
1707	20	1746	22	1785	24.1	1824	8.5	1863	44.0	1902	5.0	1941	47.5	1980	154.6
1708	10	1747	40	1786	82.9	1825	16.6	1864	47.0	1903	24.4	1942	30.6	1981	140.5
1709	8	1748	60	1787	132.0	1826	36.3	1865	30.5	1904	42.0	1943	16.3	1982	115.9
1710	3	1749	80.9	1788	130.9	1827	49.6	1866	16.3	1905	63.5	1944	9.6	1983	66.6
1711	0	1750	83.4	1789	118.1	1828	64.2	1867	7.3	1906	53.8	1945	33.2	1984	45.9
1712	0	1751	47.7	1790	89.9	1829	67.0	1868	37.6	1907	62.0	1946	92.6	1985	17.9
1713	2	1752	47.8	1791	66.6	1830	70.9	1869	74.0	1908	48.5	1947	151.6	1986	13.4
1714	11	1753	30.7	1792	60.0	1831	47.8	1870	139.0	1909	43.9	1948	136.3	1987	29.4
1715	27	1754	12.2	1793	46.9	1832	27.5	1871	111.2	1910	18.6	1949	134.7	1988	100.2
1716	47	1755	9.6	1794	41.0	1833	8.5	1872	101.6	1911	5.7	1950	83.9	1989	157.6
1717	63	1756	10.2	1795	21.3	1834	13.2	1873	66.2	1912	3.6	1951	69.4	1990	142.6
1718	60	1757	32.4	1796	16.0	1835	56.9	1874	44.7	1913	1.4	1952	31.5	1991	145.7
1719	39	1758	47.6	1797	6.4	1836	121.5	1875	17.0	1914	9.6	1953	13.9	1992	94.3
1720	28	1759	54.0	1798	4.1	1837	138.3	1876	11.3	1915	47.4	1954	4.4	1993	54.6
1721	26	1760	62.9	1799	6.8	1838	103.2	1877	12.4	1916	57.1	1955	38.0	1994	29.9
1722	22	1761	85.9	1800	14.5	1839	85.7	1878	3.4	1917	103.9	1956	141.7	1995	17.5
1723	11	1762	61.2	1801	34.0	1840	64.6	1879	6.0	1918	80.6	1957	190.2	1996	8.6
1724	21	1763	45.1	1802	45.0	1841	36.7	1880	32.3	1919	63.6	1958	184.8	1997	21.55
1725	40	1764	36.4	1803	43.1	1842	24.2	1881	54.3	1920	37.6	1959	159.0	1998	64.3
1726	78	1765	20.9	1804	47.5	1843	10.7	1882	59.7	1921	26.1	1960	112.3	1999	93.3
1727	122	1766	11.4	1805	42.2	1844	15.0	1883	63.7	1922	14.2	1961	53.9	2000	119.6
1728	103	1767	37.8	1806	28.1	1845	40.1	1884	63.5	1923	5.8	1962	37.6	2001	111.0
1729	73	1768	69.8	1807	10.1	1846	61.5	1885	52.2	1924	16.7	1963	27.9	2002	104.0
1730	47	1769	106.1	1808	8.1	1847	98.5	1886	25.4	1925	44.3	1964	10.2	2003	63.7
1731	35	1770	100.8	1809	2.5	1848	124.7	1887	13.1	1926	63.9	1965	15.1	2004	40.4
1732	11	1771	81.6	1810	0.0	1849	96.3	1888	6.8	1927	69.0	1966	47.0	2005	29.8
1733	5	1772	66.5	1811	1.4	1850	66.6	1889	6.3	1928	77.8	1967	93.8	2006	15.2
1734	16	1773	34.8	1812	5.0	1851	64.5	1890	7.1	1929	64.9	1968	105.9	2007	7.5
1735	34	1774	30.6	1813	12.2	1852	54.1	1891	35.6	1930	35.7	1969	105.5	-	-
1736	70	1775	7.0	1814	13.9	1853	39.0	1892	73.0	1931	21.2	1970	104.5	-	-
1737	81	1776	19.8	1815	35.4	1854	20.6	1893	85.1	1932	11.1	1971	66.6	-	-
1738	111	1777	92.5	1816	45.8	1855	6.7	1894	78.0	1933	5.7	1972	68.9	-	-

Tabell A.3: *Årlig registrering av solflekker 1700-2007.*

Referanser

[Brown og Churchill 2003] Brown, J og Churchill, R. (2003) "Complex Variables and Applications", 7th edition, McGraw-Hill.

[Brockwell og Davis 2002] Brockwell, P og R. Davis (2002). "Introduction to Time Series and Forecasting", 2nd edition. Springer-Verlag, New York, Inc.

[Casella og Berger 2002] Casella, G og R. Berger (2002). "Statistical Inference", 2nd edition, Thomson Learning.

[Dobson 2002] Dobson, a. (2002). "An Introduction to Generalized Linear Models", 2nd edition. Chapman & Hall/CRC.

[Edgeworth 1888] Edgeworth, F. (1888). "On a New Method of Reducing Observations Relating to Several Quantities". Philosophical Magazine, 25, 184-191.

[Efron 1979] Efron, B. (1979). "Bootstrap methods: Another look at the Jackknife". The annals of Statistics, vol 7, nr 1.

[Galton 1889] Galton, F. (1889). "Natural Inheritance". London and New York MacMillan and Co.

[Hao og Naiman 2007] Hao, L. og D. Naiman (2007). "Quantile regression". Sara Miller McCune, SAGE Publications, Inc.

[Johnson m. fl 1994] Johnson, N. L; S. Kotz, and N. Balakrishnan (1994). "Continuous Univariate Distributions", Volume 1, 2nd edition, Wiley and sons.

[Koenker og Bassett 1978] Koenker, R. og G. Bassett (1978). "Regression Quantiles". Econometrica, 46, 33-50.

[Koenker 2005] Koenker, R. (2005). "Quantile regression". Cambridge, UK: Cambridge University Press.

[Koenker og Xiao 2006] Koenker, R og Z. Xiao (2006). "Quantile Autoregression", Journal of the American Statistical Association, vol 101, nr 475, sept 2006, side 980-990, og etterfølgende diskusjon på side 991-1006.

[Mosteller og Tukey 1977] Mosteller, F. og J. Tukey (1977). "Data Analysis and Regression: A Second Course in Statistics". Reading, MA: Addison-Wesley.

[Peng og Huang 2008] Peng, L. og Y. Huang (2008). "Survival Analysis With Quantile Regression Models", Journal of the American Statistical Association, vol 103, nr 482, juni 2008, side 637-649.

[Walpole m.fl 2002] Walpole, R; R. Myers; S. Myers; K. Ye (2002). "Probability & Statistics for Engineers & Scientists", 7th edition. New Jersey, Prentice-Hall, Inc.

[Wasserman 2006] Wasserman, L. (2006). "All of Nonparametric Statistics", Springer Texts in Statistics.